RECLAIMING ENLIGHTENMENT

_on_the_logical_foundations_of_the_rule_of_law in_a_legitimate_algocracy_

evan iatrou

RECLAIMING ENLIGHTENMENT|
on the logical foundations of the rule of law in a legitimate algocracy|

MSc Thesis *(Afstudeerscriptie)*

written by

Evan (Evangelos) Iatrou

under the supervision of **Dr. Katrin Schulz** and **Dr. Vladislava Stoyanova,** and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
October 12, 2023	Dr. Ekaterina Shutova
	Prof. Davide Grossi
	Dr. Marijn Sax

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

To my pappou, Evangelos Theodosiou, for making me fox,

to DR. (<3) Margarita Amaxopoulou for being the NESS condition of my career,

and to Tomek Klochowicz for being both the greatest mentor and friend that I could ask for during this 3-year trip.

Thank you.

COPYRIGHTS

• **COVER:** Forever indebted to artist Edmon de Haro that gave me permission to use his artpiece as my cover. It is the cover of Kissinger's 2018 Atlantic article "*How the Enlightenment ends*" that constitutes the main premiss of my Thesis. According to the Atlantic's content reproduction policy, in order to use an image, I have to get permission by the image provider listed underneath the image's caption as I did.

• **INTRODUCTION:** (α) The picture of the Foreign Ministers of the Council of Europe's (CoE's) founding members signing the Statute of the CoE (1949) is property of the CoE. Permission for use is granted when the use concerns "*information and education on the Organisation's work*".; (β) The map of the CoE member states was designed *via* mapchart.net. I used it under the CC BY-SA 4.0 license by crediting the creators and providing a link to the source.

• **CHAPTER I:** All Figures were designed using LaTeX's `tikz package`.

• **CHAPTER II:** (α) The cover picture was designed by digital artist & old friend NICO MAVRIDI (INSTA: `@nima_draws`); (β) The Figures from Nitta and Satoh 2020 were used under the CC BY license by crediting the creators.; (γ) Figures 1, 2, & 3 were designed using LaTeX's `tikz package`.; (δ) The Figure in §3.2.1 is taken from Danziger, Levav, and Avnaim-Pesso 2011 which was published at PNAS. After the publication of a PNAS paper, third parties are allowed to use figures for non-commercial purposes (¶4 in PNAS' license to publish policy); (ε) The image of Chalkidis, Androutsopoulos, and Aletras's 2019 attention-based XAI explanation was used under the CC BY 4.0 license by crediting the creators.

• **CHAPTER III:** (α) Figures 1, 2, 3, & 4 were designed using LaTeX's `tikz package`.; (β) The clever Hans picture was taken from the "*Clever Hans*" Encyclopedia Britannica entry. According to its terms of use, the Encyclopedia allows the use of images for dissertations and other educational non-commercial purposes; (γ) The Clever Hans heatmap is taken from Lapuschkin et al. 2019. It was used under the CC BY license by crediting the creators.

• **CHAPTER IV:** All Figures were designed using LaTeX's `tikz package`.

References

Chalkidis, Ilias, Ion Androutsopoulos, and Nikolaos Aletras. 2019. "Neural legal judgment prediction in English." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4317–4323. Florence, Italy: Association for Computational Linguistics, July. https://doi.org/10.18653/v1/P19-1424.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous factors in judicial decisions." *Proceedings of the National Academy of Sciences* 108 (17): 6889–6892. https://doi.org/10.1073/pnas.1018033108.

Kissinger, Henry A. 2018. *How the Enlightenment ends: Philosophically, intellectually — in every way — human society is unprepared for the rise of artificial intelligence.* Technology. The Atlantic, June. Accessed March 1, 2023. https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. "Unmasking Clever Hans predictors and assessing what machines really learn." *Nature Communications* 10 (1): 1096. https://doi.org/10.1038/s41467-019-08987-4.

Nitta, Katsumi, and Ken Satoh. 2020. "AI applications to the law domain in Japan." *Asian Journal of Law and Society* 7 (3): 471–494. https://doi.org/10.1017/als.2020.35.

ACKNOWLEDGEMENT

I would like to begin by thanking my fellow Master of Logic students that accompanied me in this journey. Rover Samwel & Francesco Ponti for the countless hours we spent at Science Park's project rooms writing our Theses. Derek So (Wing Yi) & Tomasz Klochowicz for having the interest & patience to listen to my conundrums and for guiding me in my virgin trip to the world of philosophy. Wijnand van Woerkom for opening my eyes to the state-of-the-art advancements in the discipline of AI & Law. Alexander Lind not only for being a patient flatmate letting me occupy our shared space with the tens of books I borrowed from UvA's library, but more importantly, for helping me to clarify my thoughts at times when it was impossible to move forward. Finally, I would like to thank Tuva Bardal & Paul Talma for facilitating my Cool Logic talk where I was able to share & discuss parts of my progress.

I would further like to thank Margarita Amaxopoulou & Stergios Aidinlis for giving me the idea to work on the logical foundations of XAI that regulates human rights for the Master of Logic course *Causal Inference: Philosophical Theory and Modern Practice* taught by my Thesis' main supervisor Katrin Schulz. A project that marked the beginning of my Thesis. And a project that would not have been possible without the collaboration and guidance of Tomasz Klochowicz.

My capability of combining knowledge from all those diverse disciplines, from AI & logic to philosophy & human rights law, would not have been harvested without my Master of Logic mentor Pieter Adriaans. I can hardly imagine any other scholar with more capacity to fruitfully handle interdisciplinary knowledge with the same creativity & originality as Pieter, making him the gold standard that I intend to surpass. I would further like to thank Paul Dekker for keeping my enthusiasm about entering the world of analytic philosophy alive during the whole coronavirus ordeal, as well as for his immense intellectual & mental support. Arianna Betti for pushing me to my limits and forcing me to coherently structure the messiness of my creativity. Robert van Rooij for showing me first-handedly how formalism can shape philosophy and how philosophy can challenge the limits of logic. Sebastian de Haro Ollé & Dean McHugh for introducing me to the areas of philosophy that I worked on this Thesis with their high quality teaching. Federica Russo for guiding me in the vast literature of causality & philosophy of techno-science, as well as Stephanie Dick for introducing me to the early use of AI by the US military with its subsequent legitimacy concerns. Last but not least, I would like to thank Ronald de Haan for giving me the opportunity to explore the limits of Knowledge Representation & Reasoning as his teaching assistant for two years in a row.

My defence would not have been possible without the expertise of Davide Grossi & Marijn Sax that accepted to join the Thesis Committee and without Ekaterina Shutova that chaired it. I would especially like thank Ekaterina for allowing my Thesis to be hybrid so as to finally have an in-person graduation after all this COVID mess. I would also like to thank Lazaros Moysis and Tomasz Klochowicz for preparing me for my defense. And more than anyone, I would like to thank Sander Beckers for honouring me by reading my interpretation of his work and joining my defense to enlighten me with his feedback & questions.

I could not forget the administrative & technical support of the ILLC Office (Tuğba, Roos, and Sarah you are simply the best!), and of UvA's Library Desk & AudioVisual (AV) Department. They made my defence's hybrid form possible, bringing my first in-person graduation to life! Shout-out to Yani Riyani from the Library Desk for always being available to help with a big warm smile in her face!

Finally, I am forever indebted to Katrin and Vladislava. I know it sounds cliché, but no matter how I try to express my gratitude, my writing skills will never be good enough. The only thing I will say is that any further career milestone I achieve, it will always be owed to you. Especially to Katrin that changed me not only as a researcher, but also as a person. Oh, and Katrin, thank you for your patience! ;D

ABSTRACT

Algocracies, i.e., political orders where political power is exercised inter alia *by* or *via* algorithms, are already a reality. From algorithmic models partaking in judicial decisions and facial recognition AI in surveillance systems to police drones and spywares, more and more political orders are becoming more and more algocratic raising concerns about their *legitimacy*, the so-called *threat of algocracy*. Those concerns are usually about non-consensual data-driven *profiling* or about the implications of AI's *opacity* like the lack of informed participation. There is though a more fundamental threat to legitimacy. It is not a threat that undermines the legitimacy of a political order, but a threat that challenges what legitimacy *means* in the first place. It is a threat that challenges to end *Enlightenment*'s legitimacy paradigm where *human reason* is the means to order legitimate political orders.

Considering the above, the objective of the Thesis is to provide requirements that should comprise the foundations for engineering *algocratic* AI (henceforth ALGOAI) in order to avoid Enlightenment foretold death. I focus on two types of such requirements: (α) requirements for the *logical structure* of ALGOAI's output and architecture that influence the AI's *explanatory* power.; (β) *meta-scientific* requirements for the practice of engineering ALGOAI models with such logical requirements, especially for the practice of *logicians & formal philosophers*. I include *trans*disciplinary requirements, i.e., requirements that *transcend* scientific disciplinary practice like societal, political, ethical, and legal *values*. From all those values, the *rule of law* reigns supreme both in terms of *ontological priority* as well as *universality* across political orders. Ergo, I focus on ALGOAI that is used *as* or *by judicial* authorities, the quintessential paradigm of AI that threatens Enlightenment's legitimacy paradigm. Many of the results can be generalised to other types of ALGOAI as well.

Regarding (α), I start in §1 by arguing that (legal) ALGOAI engineering practice is centered around *evaluative judgements* about specific legitimacy *values* (e.g., the rule of law, human rights, democracy) *contra* other disciplinary practices where the main practice consists of *factual* judgments. I further argue what type of *justification* those evaluative judgments have in the Enlightenment legitimacy paradigm, why this type of justification is now threatened, and what should (legal) ALGOAI engineers do in order to respond to this threat. My proposal centers around the *logical structure* of those justifications. I ground my proposal on a generalisation of *Benacerraf's dilemma* from the philosophy of mathematics to meta-ethics, what I name *Benacerraf's curse*. It is essentially a problem of *ambiguity of meaning*. In §3, I argue why and how *conceptual re-engineering*, and in particular Carnap's method of *explication*, can be used to engineer legal ALGOAI models that satisfy the foregoing requirements. In particular, explication should be used to re-engineer concepts of *judicial reasoning* used in the actual legal practice. In §4, I provide a toy example of how explication can be applied to judicial *causal reasoning* in order to contribute to the engineering of legal ALGOAI intended to be used by the European Court of Human Rights (ECtHR). I focus on the so-called *NESS* and *but-for causal justifications*. My goal is *not* to provide a full-fledged account of an explicated concept of causal justification but to show how explication can be performed.

Regarding (β), I contextualise my proposal in the context of *philosophy of interdisciplinarity*, the nascent evolutionary stage of *philosophy of science*. More precisely, in §2, I argue about *which* disciplines should collaborate and *how* in order to engineer legal ALGOAI models based on the requirements I introduced in §1. I put emphasis on the role that *logicians & formal philosophers* should have in an ALGOAI engineering team. *Contra* traditional philosophy of science, philosophy of interdisciplinarity emphasizes the need to engineer ALGOAI based on *transdisciplinary meta-scientific* requirements. Considering this, I provide such meta-scientific transdisciplinary requirements like the role of legal ALGOAI engineers in the new system of checks and balances that characterises algocratic orders. ALGOAI engineers are no longer mere engineers, but they are *political actors* that (co-)exercise *political power*. Once more I focus on the normative contribution of logicians & formal philosophers to those transdisciplinary requirements. Finally, I contextualise those transdisciplinary requirements in the context of the emerging 5th industrial revolution and the new social order predicated on it, the so-called SOCIETY 5.0.

Contents

NOTE TO THE READER

• Whenever I am referring to a section of a CHAPTER Y (e.g., section 4.1.2 from CHAPTER II) in another CHAPTER X (e.g., in CHAPTER I), I am using the index Y in my reference (e.g., §II.4.1.2). Whenever I am referring to a section of the *same* chapter (e.g., referring to section 2.2 of CHAPTER I in CHAPTER I itself), I am not using the chapter index in my reference (e.g., using §2.2, not §I.2.2).
• Whenever I use SMALL CAPS to refer to a specific legal provision, I will be referring to articles of the European Convention of Human Rights.
• *Endnotes* are marked by lowercase Roman numerals, while *footnotes* by Arabic numerals. Endnotes should be construed as part of the APPENDIX, while footnotes as part of the main text.
• Figures', endnotes', and footnotes' numberings are reset at the beginning of each CHAPTER.
• Instead of *she/her*, whenever I want to stay neutral regarding one's gender, I am using *they/them*.

ABBREVIATIONS

INSTITUTIONS, ORGANISATIONS, AND THE LIKE

AFP	Agence France-Presse
CAI	Committee on Artificial Intelligence
CEPEJ	*fr*: Commission Européenne Pour l'Efficacité de la Justice *eng*: European Commission for the Efficiency of Justice
CERI	Centre for Educational Research and Innovation
CJEU	Court of Justice of the European Union (EU)
CoE	Council of Europe *fr.* Conseil de l'Europe (CdE)
the Court	European Court of Human Rights *aka:* EC(t)HR *fr.* Cour européenne des Droits de l'Homme (CEDH, CrEDH, ou CourEDH)
CSO	Civil Society Organisation
DW	Deutsche Welle
EC	European Commission
EC(t)HR	European Court of Human Rights, Some use "*ECHR*" as an abbreviation for the European Convention of Human Rights *aka:* the Court (capitalised "*C*"), the Strasbourg Court *fr.* Cour européenne des Droits de l'Homme (CEDH, CrEDH, ou CourEDH)
EIU	Economists Intelligence Unit
EP	European Praliament
EU	European Union
eu-LISA	EU agency for the operational management of large-scale IT systems in the area of freedom, security and justice
Eurojust	European Union Agency for Criminal Justice Cooperation)
FGCS	Fifth Generation Computer Systems project
GC	Grand Chamber
ICJ	International Court of Justice
IIEA	Institute of International and European Affairs
ILC	International Law Commission
JURI	EP's Committee on Legal Affairs
LIBE	EP's Committee on Civil Liberties, Justice and Home Affairs

MEP	Member of the European Parliament
MP	Members of the Parliament
MSI-AUT	Committee of experts on Human Rights dimensions of automated data processing and different forms of artificial intelligence
NATO (*fr.* OTAN)	North Atlantic Treaty Organization (*fr.* Organisation du Traité de l'Atlantique Nord)
NGO	Non-Governmental Organisation
OECD	Organisation for Economic Co-operation and Development
PACE	Parliamentary Assembly of the Council of Europe
PEGA	Committee of inquiry to investigate the use of Pegasus and equivalent surveillance spyware
SDG	Sustainable Development Goal *aka:* Global Goals
UN	United Nations
UN ILC	United Nations International Law Commission
UNESCO	United Nations Educational, Scientific and Cultural Organization
Venice Commission	European Commission for Democracy through Law
WJP	World Justice Project

LEGAL PROVISIONS

(ILC) ARSIWA	ILC Articles on the Responsibility of States for Internationally Wrongful Acts
the Convention	the European Convention of Human Rights
EUDPR	Regulation (EU) 2018/1725
GDPR	General Data Protection Regulation
LED	Law Enforcement Directive
LOAC	Law of Armed Conflict *aka:* International Humanitarian Law
MPC	Model Penal Code
UDHR	Universal Declaration of Human Rights

AI & COMPUTER SCIENCE

ADM	Algorithmic Decision-Making system
CADx	Computer-Aided Diagnosis
CBR	Case-Based Reasoning
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CSO	Civil Society Organisation
GOFAI	Good Old-Fashioned AI
GPT	Generative Pre-trained Transformer
HART	Harm Assessment Risk Tool
ICT	Information and Communications Technology
LAW	Lethal Autonomous Weapon
ML	Machine Learning
NNs	Neural Networks
RBR	Rule-Based Reasoning
VR	Virtual Reality
XAI	Explainable Artificial Intelligence

LAW, POLITICAL SCIENCE, AND HUMANITIES

HCP	High Contracting Party *aka:* Contracting State, State Parties
HUDOC	Database of the ECtHR's case-law: https://hudoc.echr.coe.int/eng
MENA	Middle East and North Africa area
MEP	Member of the European Parliament
MP	Member of the Parliament
R&D	Research & Development
PM	Prime Minister
SMEs	Small & Medium Enterprises
WW	World War

PHILOSOPHY & LOGIC

CMS	Classical Model of Science
CROSSDI	Cross-disciplinarity/cross-disciplinary
DI	Disciplinarity/disciplinary
HP explication of causation	(Joseph Y.) Halpern's & (Judea) Pearl's explication of causation
ID	Interdisciplinarity/interdisciplinary
INUS condition	Insufficient, but Necessary part of an Unnecessary but Sufficient condition
METADI	Meta-disciplinarity/meta-disciplinary
MULTIDI	Multi-disciplinarity/multi-disciplinary
NESS	Necessary Element of a Sufficient Set
PhID	Philosophy of Interdisciplinarity
PhilSci	Philosophy of Science
PLURIDI	Pluri-disciplinarity/pluri-disciplinary
RE	Reflective Equilibrium
SEM	Structural Equation Models
SCM	Structural Causal Models
SEP	Stanford Encyclopedia of Philosophy
TRANSDI	Transdisciplinarity/transdisciplinary

RECLAIMING ENLIGHTENMENT|

on the logical foundations of the rule of law in a democratic algocracy|

INTRODUCTION

A death foretold

> "*The Enlightenment started with essentially philosophical insights spread by a new technology. Our period is moving in the opposite direction. It has generated a potentially dominating technology in search of a guiding philosophy.*"

The foregoing quote is from an 2018 article Henry Kissinger ominously named "*How the Enlightenment ends: Philosophically, intellectually — in every way — human society is unprepared for the rise of artificial intelligence*" (Kissinger 2018; *see also* Kissinger, Schmidt, and Huttenlocher 2019, 2021). Kissinger has been an architect of the current world order whose "*world-ordering*" activity was an exemplar of realpolitik (Merlini 2023), an approach of engaging in international relations by prioritising pragmatic over idealistic *desiderata* (Brown, McLean, and McMillan 2018). When such an adherent of *pragmatism* worries about the displacement of philosophy from technology as a guide of world ordering, it is a telling sign that either they have had a change of heart or that they are no longer mentally fit or that the world is in the midst of radical changes. And 2018 Kissinger was neither regretful nor mentally incompetent. Many of Kissinger exact worries may end up being misplaced. They may be a technophobic reflex; the fear of a new, uncharted territory. Still, something fundamental *is* changing. And we should make sure that we are guarded from any implications of those changes.

I *do* share Kissinger's concern. In what Kissinger construes as the Enlightenment era, what *legitimises* a political order is the use of *human reason* to determine the *content* of the *values* that order is expected to uphold (§I.2.3; *cf.* Kissinger 2018). It does not matter if it is a layman discourse, a debate among antagonising politicians, a heated peer-reviewed exchange among academics or the drafting of a judgment by the judges of the highest court. At the end of the day, knowingly or unknowingly, when one argues how those values should be, then they are engaging in a philosophical discourse grounded on human reason and whose foundations were laid down during Enlightenment (*ibid.*). However, the emergence of AI shifts the content of those values in ways *not* pre-decided by human reason. The more AI is involved in our lives, from judicial decisions to national defense, the more those values shift content independently of any philosophically-based justifications (§I.3.2.1). As Kissinger remarks in the quoted passage, technology no longer spreads the philosophical insights that give to values their human-reason-determined content (henceforth simply "*rationally determined*"), but it is technology itself that determines that content disseminating its own new "*guiding philosophy*". Enlightenment's legitimacy paradigm is *disrupted*.

The root of this disruption is summed up in the following remark from Kissinger's article:

> "*Through all human history, civilizations have created ways to explain the world around them—in the Middle Ages, religion; in the Enlightenment, reason; in the 19th century, history; in the 20th century, ideology. The most difficult yet important question about the world into which we are headed is this: What will become of human consciousness if its own explanatory power is surpassed by AI, and societies are no longer able to* ***interpret*** *the world they inhabit in terms that are* ***meaningful to them****?*"[1]

When Kissinger refers to "*explanatory power*", he does not mean the ability of AI to explain its output, what is referred to as the *opacity* (or *black-box*) concern (§I.3.2; §II.4.2.2). He means something more fundamental than that: the possibility that the way that AI explains ("*interprets*") the world is superior to any explanation

[1] Kissinger 2018, emphasis added.

(*interpretation*) of the world produced by human reason alone. As he later put it in a book that he co-authored with two renowned AI experts (Kissinger, Schmidt, and Huttenlocher 2021, pp.49-50, emphasis added):

> "*...the central Enlightenment premise of a knowable world being unearthed, step-by-step, by human minds has persisted. Until now. Throughout three centuries of discovery and exploration, humans have* ***interpreted the world*** *as Kant predicted they would according to the structure of* ***their own minds****. But as humans began to approach the limits of their cognitive capacity, they became willing to enlist machines — computers — to augment their thinking in order to* ***transcend those limitations.***"

Consequently, if AI can eventually identify and realise values "*superior*" to those that human reason can conceive, whatever one means by "*superior*", then humanity is faced with what I name *Ismene's dilemma* (§I.3.2.1.2): should we allow AI to re-order our world based on values that we can not *rationally* interpret but we still *believe* to be superior to ours or should we continue using AI as means to realise *only* our own *rationally* determined ends? If one chooses the former, then they have to argue why this *faith* to a superior uninterpretable by reason authority is different from the pre-Enlightenment paradigm of a God-driven political order. If one chooses the latter, then they have to argue why should humanity reject and not embrace a technology that outperforms its bounded epistemic abilities.

Ismene's dilemma is not a new story. It is a resurgence of the old *philosopher king* debate that re-emerged with the form of *epistocracy*: those that have an epistemic privilege over the others, those should rule in virtue of that privilege (§I.2.8). And for the advocates of a post-Enlightenment concept of legitimacy, contemporary state-of-the-art *artificial* intelligence has an epistemic privilege over *human* intelligence and ergo it can *legitimately* exercise political power like partaking in judicial decision-making. Danaher 2016 named the epistocratic challenges to the legitimacy of contemporary political orders induced by algorithmic decision-making systems (ADMs)[2] that exercise political power as the *threat of algocracy*, where *algocracy* is any governance system that accommodates ADMs that exercise power (*cf.* §I.3, ¶1). I will call any such ADM as *algocratic ADM* and any such AI ADM as *algocratic AI* (henceforth "*algocratic*" will be abbreviated as "*Algo*"). Furthermore, I construe as *AlgoAI engineers* all the experts that participate in the *designing*, *building*, and *analysis* of AlgoAI models (more on the three phases of engineering on §II.4.1.1).

My approach to the threat of algocracy differs from that of Danaher 2016. I do not dismiss Danaher's legitimacy concerns. Quite the opposite actually. As I argue in §I.3.2, Danaher hit the nail on the head by raising legitimacy concerns that constitute central premisses of contemporary state-of-the-art AI research like engineering AI that *explains* its output (the so-called explainable AI or XAI) (§II.4.2.2). The legitimacy concern that I introduce though is an additional threat to legitimacy albeit a more fundamental one. While Danaher warns about how AlgoAI can make a political order *il*legitimate, motivated by Kissinger's remarks, I go one step further arguing how AlgoAI *disrupts* the current legitimacy paradigm rooted in Enlightenment by challenging the *meaning* of legitimacy. It is not a concern about whether a political order is illegitimate or not but about what it *means* to be legitimate in the first place.

The quintessential example of AlgoAI that threatens Enlightenment's legitimacy paradigm is AlgoAI that is used to *interpret & apply* the law (henceforth *legal AlgoAI*) (§I.3.2.1, ¶2). The respective research field is that of *AI & Law* (Bench-Capon et al. 2012). Taking this into consideration, in order to introduce the threat of algocracy, I will use as an example a particular case of legal AlgoAI. I choose legal AlgoAI that is intended to be used by the *European Court of Human Rights* (*abbr: ECtHR* or simply *the Court* (capitalised "C"); *fr: Cour Européenne des Droits de l'Homme* (*abbr: CEDH*, *CrEDH*, ou *CourEDH*)) either as a substitute of judges (what is called *replacement AI*)[3] or as a supportive tool to human judges (*supportive AI*).[3] Apart from my personal interest in European politics, I chose the ECtHR *contra* other judicial authorities mainly for the following three reasons. Firstly, the ECtHR's judgements are *binding* for 46 European countries of more than 700 million people widening the range of my argument's applicability. In addition to that, many of the values the ECtHR is expected to uphold so as to be legitimate are the *bear minimum* of values that more or less all post-WWII political orders are expected to uphold so as to be legitimate (§I.2) widening even further my argument's applicability. Secondly, the ECtHR's case-law is publicly available[4] in multiple languages including English allowing me to access a vast amount of documents that have been processed for years by academics from a diverge range of disciplines, from AI & Law (e.g., Moreira 2022; Medvedeva et al. 2020; Kaur and Božzić 2020) to political

[2] Abbreviation taken from MSI-AUT 2019, §2.1. Danaher 2016 uses "*algorithms*" instead of "*ADMs*" (p.247), but I deem it too broad and imprecise.

[3] Winter, Hollman, and Manheim 2023, p.188; *see also* §I.2.5, ¶8.

[4] *See* its database HUDOC as well as the diverse documents provided by the ECtHR's organs to help understanding how the ECtHR makes its judgements. E.g., the ECtHR Registry's guides on specific Convention articles or the ECtHR Press Service's factsheets on specific types of cases (e.g., human trafficking, prisoners' voting rights, protection of reputation).

science (e.g., Stiles 2006; Shattock 2022) and jurisprudence[5] (e.g., Letwin 2021; Turton 2020). Finally, while due to its importance and accessibility the ECtHR's case-law has spawned a rich academic literature especially in Europe, there is a sparsity of representation in the literature of *analytic* philosophy. The latter is quite common in the Anglo-American legal tradition leading to significant advancements not only in the philosophy of law, but also in multiple other areas of analytic philosophy, in logic, AI, as well as in the actual legal practice (§IV). This Thesis aspires to make a small step towards bridging the gap between the ECtHR practice and analytic philosophy in the footsteps of other works like Letsas 2007.

More precisely, in CHAPTER I, I introduce the threat of algocracy in the context of the 5^{th} *Industrial Revolution* (INDUSTRY 5.0) and I argue which *logical* requirements should legal ALGOAI models have in order to avoid Enlightenment's foretold death. In CHAPTER II, I argue how the ALGOAI engineers, a team of experts from the disciplines of AI, law, logic, & formal philosophy, should cooperate in order to satisfy the logical requirements I introduced in CHAPTER I. I do so in the context of *philosophy of interdisciplinarity* (*PhID*), the nascent evolutionary stage of philosophy of science (PhiSci). In CHAPTER III, I argue why & how Carnap's *conceptual engineering* method of *explication* can be used to engineer ALGOAI satisfying the legitimacy requirements of CHAPTERS I & II. Finally, in CHAPTER IV, I showcase how the account of explication laid out in CHAPTER III can be applied in a particular example. That example is the explication of the concept of *causal justification* as it is used in the ECtHR's case-law to *justify* its judgements, an explication whose purpose is to be used by ALGOAI engineers against the threat of algocracy.

In the rest of the INTRODUCTION, I introduce the European Court of Human Rights and the *three legitimacy pillars* of the post-WWII European order that the ECtHR is expected to protect: democracy, human rights, & the rule of law.

[5] "*jurisprudence*" is an alternative term for the *philosophy of law* (Leiter and Sevel 2022).

Legitimacy in the European order

> "*I wish to speak to you today about the tragedy of Europe.*"
>
> *On a Council of Europe*, Winston Churchill
> Zurich University, 19 September 1946

05 May 1949, St James's Palace, London: The Foreign Ministers of the CoE's founding members (Belgium, Denmark, France, Ireland, Italy, Luxembourg, the Netherlands, Norway, Sweden, UK) signing the Statute of the CoE, also known as the Treaty of London (1949), that marked the birth of Europe's oldest political body.

Democracy, human rights, and the rule of law have been the foundational values of the post-WWII European order's legitimacy. They are not mere normative "*Ivory-towerish*" requirements set by philosophers in a Sanford Encyclopedia of Philosophy (SEP) entry.[6] They are requirements that have *explicitly* fleshed out in national and international legal provisions[7] drafted and ratified throughout those 70+ years by European political authorities people. The quintessential archetype of this legitimacy paradigm is the Council of Europe (CoE) (*fr. Conseil de l'Europe (CdE)*), the oldest European political body that was established on 05 May 1949, right after the end of WWII (BBC 2010), to ensure that the world order will not be decayed again to its rotten wartime state (Weiß 2017, §B; *see also* Nussberger 2020, p.3; Holm 2023, p.19). A political body in which 47 out of the current 51 European countries have participated at some point by ceding national sovereignty in order to ensure that the European order is founded upon *CoE's three pillars*: democracy, human rights, and the rule of law (Weiß 2017, ¶¶1.39,1.49,1.54-1.55,1.62-1.64; Bond 2012, §10; CDL-AD(2011)003rev, p.1; PACE 1992, pp.230-231).

A similar initiative took place a few years earlier at an international level with the establishment of the United Nations (UN) on 24 October 1945. European countries though were uneased by the fact that the UN's Universal Declaration of Human Rights (UDHR) would end up being a non-legally binding document. On May 1948, responding to Winston Churchill's call,[8] a "*frustrated*" Netherlands[9] held "*The Congress of Europe*", a conference in which politicians, members of European parliaments and governments, representatives from employers' organisations and trade unions, journalists, and intellectuals from seventeen different European states (including the former Axis powers' states) proposed *inter alia* the signing of a human rights document by European states that would be applied by a European supreme court (Weiß 2017, §B; Oomen 2016, p.411-413; Bond 2012, §6; CVCE, n.d.). This time, the legal provisions protecting human rights would be legally binding.

One year later, at St Jame's Palace, London, the Foreign Ministers of Belgium, Denmark, France, Ireland, Italy, Luxembourg, the Netherlands, Norway, Sweden, and the UK, will sign the Statute of the CoE, also known as the Treaty of London (1949), marking the beginning of the CoE. Greece followed three months later, with Türkiye, Western Germany, and Iceland joining the coming year (Bond 2012, §6). The Treaty of London erected CoE's three pillars. In Article 3, the member states commit that they will abide by "*the principles of the rule of law*" and that they will protect human rights. According to the Statute's preamble, the concepts of rule of law and human rights "*form the basis*" of a "*genuine democracy*". Under this premiss, Article 1 of the Statute establishes that the CoE will pursue the introduction of "*organs*" that will contribute to the "*maintenance*" and

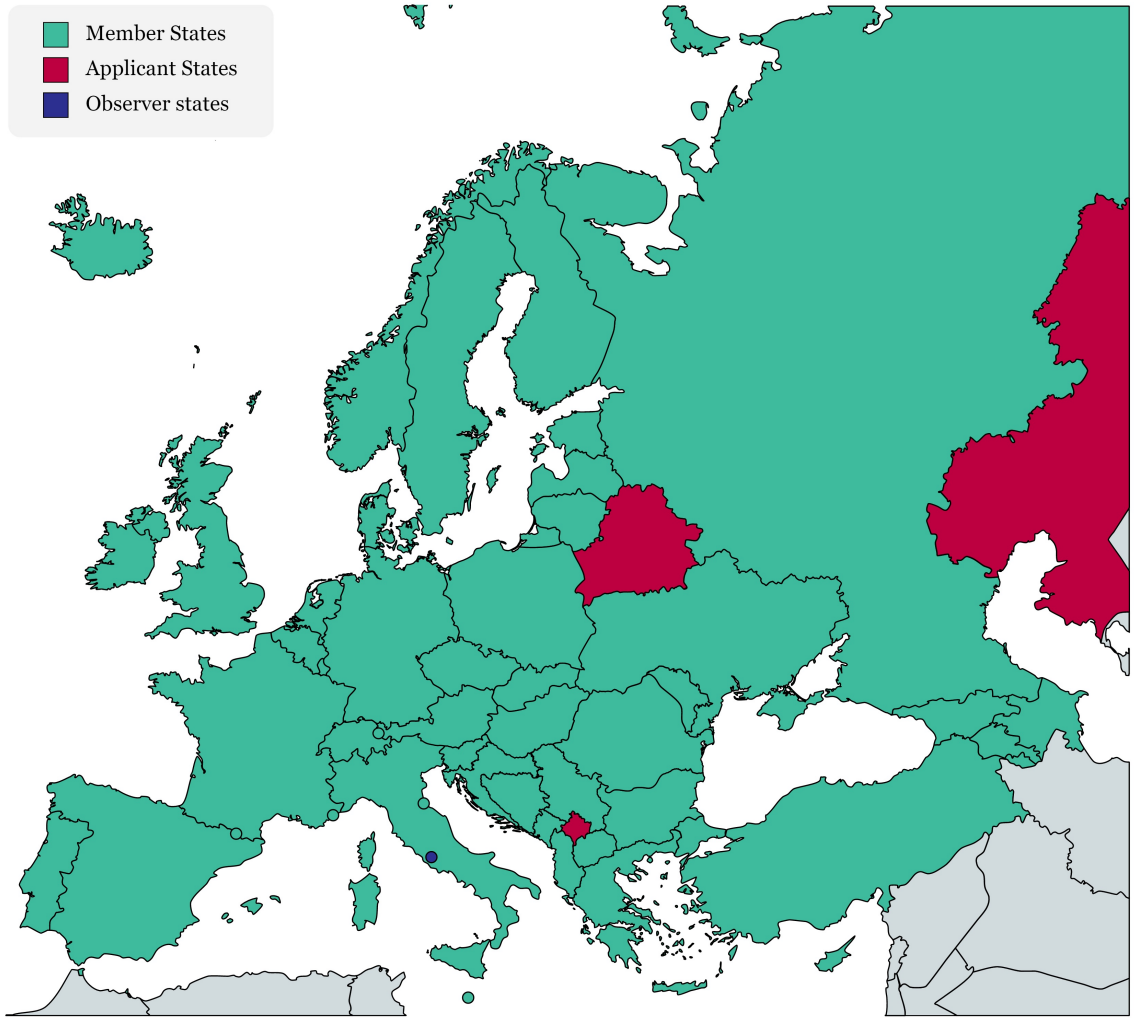
[6] In the SEP entry on the *rule of law*, Waldron argues that *rule of law*, *democracy*, and *human rights* are construed by philosophers as standard foundational values of the *liberal* political morality (Waldron 2020, §1). As we will see in §§I.1-I.2, liberal morality is the morality that characterises the post-WWII European order's *legitimacy* (*see also* Kissinger 2014; Huntington 2011, §3). It is also the quintessential political morality of *Enlightenment* (§I.2.3, ¶4).

[7] I construe "*legal provision*" as an umbrella term that refers to any type of *authoritative* text (e.g., laws, treaties, international human rights instruments) whose authoritative content is about regulating the behaviour of a group of agents. That group of agents constitutes the *jurisdiction* of the legal provision. I base this construal on Governatori, Rotolo, and Sartor 2021, p.664.

[8] *See* Churchill's landmark "*On a Coucil of Europe*" speech delivered at Zurich University on 19 September 1946: http://aei.pitt.edu/14362/ (Accessed February 25, 2023).

[9] Oomen 2016, p.411.

"*further realisation*" of human rights. And indeed, the CoE in its more than 70 years of operation has established multiple organs that ensure that its member states not only maintain and further realise human rights (e.g., the *European Court of Human Rights* (*ECtHR*), the *Committee for the Prevention of Torture* (*CPT*), or the AI-oriented *MSI-AUT* committee[10]), but they also maintain and further realise the principles of the rule of law (e.g., the *European Commission for Democracy through Law* henceforth the *Venice Commission*, the *European Commission for the Efficiency of Justice* henceforth the *CEPEJ* from the french *Commission Européenne Pour l'Efficacité de la Justice*), as well as to maintain and further realise healthy democracies (e.g., the *Parliamentary Assembly* (*PACE*), *the Congress of local and regional authorities*).



MAP DESIGNED USING mapchart.net

This is the map of the CoE's member states until the decision to expel Russia on 15 March 2022 (CoE's Newsroom 2022). From left to right, the applicant states are Kosovo,[11] Belarus, Kazakhstan. In order to be accepted, the applicant state States need to adjust their legislation so as to meet the legal standards set by the ECtHR (e.g., abolishing the death penalty).[12] The observer state is the Holy See. The map does not include the European states' overseas territories (e.g, Denmark's Faroe Islands and Greenland) and the non-European observer states (Canada, Japan, Mexico, USA; for more *see* https://www.coe.int/en/web/der/observer-states (accessed 02 February, 2023)). Also, some of the map's non-European states (e.g., Morocco, Israel) have relations with the CoE that are not contained in the legend (*see* respectively link1 & link2 (accessed 01 July, 2023)).

The CoE's organ that has achieved its most important milestones is the *ECtHR*. It is the court that the participants of the Congress of Europe envisaged in 1948. The court that supervises the *European Convention of Human Rights* (henceforth *the Convention*; *fr: Convention européenne des Droits de l'Hommes*), a legally binding

[10] *Committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence.*
[11] Note that until August 2023 Kosovo has not been recognised as a state by all CoE member states (e.g., Spain, Romania, Greece).

document that protects the three pillars across all CoE's states.[12] Until the decision of 15 March 2022 to expel Russia from the CoE due to the invasion of Ukraine, all but 4 European states[i] had ratified the Convention willingly committing to protecting human rights for more than 700 million Europeans, from performing major changes to their legislation[13] like the abolition of the death penalty (Nussberger 2020, p.24) to paying pecuniary damages to the individuals whose human rights they violated (ARTICLE 41; *cf.* Nussberger 2020, pp.161-164).

Considering actuality though, the idealistic picture I am painting is not reflective of how European states ground their legitimacy. The massive and continuously incrementing number of cases submitted to the Court every year, reaching the number of 77.400 cases as of March 2023 (ECtHR's Press Unit 2023, p.1), show that it is one thing pleading to protect human rights and a whole nother story actually doing it. Likewise for the concepts of democracy and the rule of law. After all, the three pillars are inextricably intertwined (§I.2.4, ¶2). According to WJP's 2022 Rule of Law Index®, Germany, Finland, Norway, Austria, Slovak Republic, and Portugal scored lower in the overall index than the year before. At the same time, according to EIU's 2022 Democracy Index, 22 CoE member states have been classified as *flawed democracies*, 5 CoE member states have been classified as *hybrid regimes*, and 1 as *authoritarian*.[ii]

Still, despite any divergence from the three pillars, democracy, human rights, and the rule of law remain the *normative* requirements for an authority's legitimacy in the contemporary European political order. Ergo, for AI to partake in the exercise of power, it has to be grounded on those pillars. It should at the very least not contradict them and at the very best reinforce them.

Endnotes

i. Belarus, Holy See (aka Vatican City), Kazakhstan, Kosovo are the remaining 4 states. Note that until August 2023 Kosovo has not been recognised as a state by all CoE member states (e.g., Cyprus, Slovakia, Spain).

For reasons of completeness, the exhaustive list of the rest European countries is: Albania, Andorra, Armenia, Austria, Azerbaijan, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Monaco, Montenegro, Netherlands, North Macedonia, Norway, Poland, Portugal, Republic of Moldova, Romania, Russian Federation, San Marino, Serbia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Türkiye, Ukraine, United Kingdom.

ii. FLAWED DEMOCRACIES: These countries also have free and fair elections and, even if there are problems (such as infringements on media freedom), basic civil liberties are respected. However, there are significant weaknesses in other aspects of democracy, including problems in governance, an underdeveloped political culture and low levels of political participation.

HYBRID REGIMES: Elections have substantial irregularities that often prevent them from being both free and fair. Government pressure on opposition parties and candidates may be common. Serious weaknesses are more prevalent than in flawed democracies—in political culture, functioning of government and political participation. Corruption tends to be widespread and the rule of law is weak. Civil society is weak. Typically, there is harassment of and pressure on journalists, and the judiciary is not independent.

AUTHORITARIAN REGIMES: In these states, state political pluralism is absent or heavily circumscribed. Many countries in this category are outright dictatorships. Some formal institutions of democracy may exist, but these have little substance. Elections, if they do occur, are not free and fair. There is disregard for abuses and infringements of civil liberties. Media are typically state-owned or controlled by groups connected to the ruling regime. There is repression of criticism of the government and pervasive censorship. There is no independent judiciary.

The definitions are taken *verbatim* from EIU 2023, §Appendix, p.67.

[12] CoE's member states are obliged to ratify the Convention within one year of their accession to the CoE: "*...accession to the Council of Europe must go together with becoming a party to the European Convention on Human Rights. It therefore considers that the ratification procedure should normally be completed within one year after accession...*" (Resolution 1031 (1994), ¶9). The member states that have ratified the Convention are called ***High Contracting Parties*** (***HCPs***).

[13] For comparative studies on the ECtHR's impact among the legal orders of the HCPs *see* Keller and Stone Sweet 2008; Cohen-Jonathan 1994. For the ECtHR's impact on the Dutch legal order *see* Alkema 1994; Danelius 1994; Kooijmans 2010; Hommes 2023. For the ECtHR's impact on the eastern European legal order *see* Letnar Černič 2018.

References

Alkema, Evert. 1994. "The effects of the European Convention of Human Rights and other international human rights instruments on the Netherlands legal order." In *The dynamics of the protection of human rights in Europe: Essays in honour of Henry G. Schermers*, edited by Rick Lawson and Matthijs de Blois, III:1–14. Martinus Nijhoff Publishers.

BBC. 2010. *Profile: The Council of Europe.* December 11, 2010. Accessed March 2, 2023. http://news.bbc.co.uk/2/hi/europe/country_profiles/4816408.stm.

Bench-Capon, Trevor, Michał Araszkiewicz, Kevin Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier, et al. 2012. "A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law." *Artificial Intelligence and Law* 20:215–319. https://doi.org/10.1007/s10506-012-9131-x.

Bond, Martyn. 2012. *The Council of Europe: Structure, history and issues in European politics.* Global Institutions Series. Routledge.

Brown, Garrett W., Iain McLean, and Alistair McMillan, eds. 2018. *A concise Oxford dictionary of politics and international relations.* 4th ed. Oxford University Press.

Centre Virtuel de la Connaissance sur l'Europe (CVCE). n.d. *Le Congrès de l'Europe à La Haye (7 au 10 mai 1948).* Université du Luxembourg, Luxembourg Centre for Contemporary and Digital History (C^2DH). Accessed March 10, 2023. https://www.cvce.eu/education/unit-content/-/unit/7b137b71-6010-4621-83b4-b0ca06a6b2cb/4b311dc0-cbe6-421d-9f9a-3bc8b1b155f6.

Churchill, Winston. 1946. *On a Council of Europe.* Zurich University, September 19, 1946. Accessed February 25, 2023. http://aei.pitt.edu/14362/.

Cohen-Jonathan, Gérard. 1994. "Les rapports la Convention européenne des Droits de l'Hommes et les autres traités conclus par les Etats Parties." In *The dynamics of the protection of human rights in Europe: Essays in honour of Henry G. Schermers*, edited by Rick Lawson and Matthijs de Blois, 3:79–112. Martinus Nijhoff Publishers.

Council of Europe's (CoE's) Newsroom. 2022. *The Russian Federation is excluded from the Council of Europe.* May 16, 2022. Accessed September 1, 2022. https://www.coe.int/en/web/portal/-/the-russian-federation-is-excluded-from-the-council-of-europe.

Danaher, John. 2016. "The threat of algocracy: Reality, resistance and accommodation." *Philosophy and Technology* 29 (3): 245–268. https://doi.org/10.1007/s13347-015-0211-1.

Danelius, Hans. 1994. "Article 3 ECHR and asylum law and practice in the Netherlands." In *The dynamics of the protection of human rights in Europe: Essays in honour of Henry G. Schermers*, edited by Rick Lawson and Matthijs de Blois, 3:113–122. Martinus Nijhoff Publishers.

ECtHR's Press Unit (Unité de la Presse). 2023. *Factsheet on pilot judgements.* March. Accessed April 28, 2023.

EIU (Economist Intelligence Unit). 2023. *Democracy Index 2022: Frontline democracy and the battle for Ukraine.* Published by the Economist Intelligence Unit (EIU). https://www.eiu.com/n/campaigns/democracy-index-2022/.

Governatori, Guido, Antonino Rotolo, and Giovanni Sartor. 2021. "Logic and the law: philosophical foundations, deontics, and defeasible reasoning." Chap. 9 in *Handbook of deontic logic and normative systems*, edited by Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, vol. 2. College Publications.

Holm, Michael. 2023. "The other American Dream: The one world order and Human Rights." Chap. 1 in *How democracy survives: Global challenges in the Anthropocene*, edited by Michael Holm and R. S. Deese, Part I: The forgotten promise of 1945, 9–28. Democratization and Autocratization Studies. Routledge.

Hommes, Weibe. 2023. "Co-creating European human rights: How the Netherlands received and shaped the European Convention on Human Rights, 1945- 2022." PhD diss., Faculty of Law (Faculteit der Rechtsgeleerdheid), Universiteit van Amsterdam.

Huntington, Samuel P. (1996) 2011. *The clash of civilisations and the remaking of world order.* Foreword by Zbigniew Brzezinski. Simon & Schuster.

Kaur, Arshdeep, and Bojan Božzić. 2020. "Convolutional neural network-based automatic prediction of judgments of the European Court of Human Rights." *27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, CEUR Workshop Proceedings* 2563:458–469.

Keller, Helen, and Alec Stone Sweet, eds. 2008. *A Europe of rights: The impact of the ECHR on national legal systems.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199535262.001.0001.

Kissinger, Henry A. 2014. *World order.* Penguin Press.

———. 2018. *How the Enlightenment ends: Philosophically, intellectually — in every way — human society is unprepared for the rise of artificial intelligence.* Technology. The Atlantic, June. Accessed March 1, 2023. https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

Kissinger, Henry A., Eric Schmidt, and Daniel Huttenlocher. 2019. *Metamorphosis: AI will bring many wonders. It may also destabilize everything from nuclear détente to human friendships. We need to think much harder about how to adapt.* Technology. The Atlantic, August. Accessed March 20, 2023. https://www.theatlantic.com/magazine/archive/2019/08/henry-kissinger-the-metamorphosis-ai/592771/.

———. 2021. *The age of AI: And our human future.* Little, Brown and Company.

Kooijmans, Tijs. 2010. "The burden of proof in confiscation cases: A comparison between the Netherlands and the United Kingdom in the light of the European Convention of Human Rights." *European Journal of Crime, Criminal Law and Criminal Justice* 18:225–237.

Leiter, Brian, and Michael Sevel. 2022. *Philosophy of law.* Online ed. Revised and updated by Jeannette L. Nolen. Encyclopedia Britannica, August 9, 2022. Accessed April 1, 2023. https://www.britannica.com/topic/philosophy-of-law.

Letnar Černič, Jernej. 2018. "Impact of the European Court of Human Rights on the rule of law in central and eastern Europe." *Hague Journal on the Rule of Law* 10 (1): 111–137. https://doi.org/10.1007/s40803-018-0074-5.

Letsas, George. 2007. *A theory of interpretation of the European Convention on Human Rights.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199203437.001.0001.

Letwin, Jeremy. 2021. "Why completeness and coherence matter for the European Court of Human Rights." *European Convention on Human Rights Law Review* 2 (1): 119–154. https://doi.org/10.1163/26663236-bja10002.

Medvedeva, Masha, Xiao Xu, Martijn Wieling, and Michel Vols. 2020. "JURI SAYS: An automatic judgement prediction system for the European Court of Human Rights." Edited by Serena Villata, Jakub Harašta, and Petr Křemen. *Legal Knowledge and Information Systems: JURIX 2020: The 33rd Annual Conference* (Brno, Czech Republic), 277–280.

Merlini, Cesare. 2023. "Kissinger and Monnet: Realpolitik and interdependence in world affairs." *Survival* 65 (1): 129–140. https://doi.org/10.1080/00396338.2023.2172860.

Moreira, Nídia Andrade. 2022. "The Compatibility of AI in Criminal System with the ECHR and ECtHR Jurisprudence." In *Progress in Artificial Intelligence*, edited by Goreti Marreiros, Bruno Martins, Ana Paiva, Bernardete Ribeiro, and Alberto Sardinha, 108–118. Cham: Springer International Publishing.

MSI-AUT (CoE's committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence), Rapporteur: Karen Yeung. 2019. *Responsibility and AI: A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework.* Council of Europe study. DGI(2019)05. Printed at the Council of Europe.

Nussberger, Angelika. 2020. *The European Court of Human Rights.* 1st ed. (online). Edited by Mark Janis, Douglas Guilfoyle, Stephan Schill, Bruno Simma, and Kimberley Trapp. Elements of International Law. Oxford University Press. https://doi.org/10.1093/law/9780198849643.001.0001.

Oomen, B. M. 2016. "A serious case of Strasbourg-bashing? An evaluation of the debates on the legitimacy of the European Court of Human Rights in the Netherlands." *The International Journal of Human Rights* 20 (3): 407–425. https://doi.org/10.1080/13642987.2015.1100927.

PACE (Parliamentary Assembly of the Council of Europe). 1992. "The geographical enlargement of the Council of Europe." *Human Rights Law Journal* 13:230–236.

Shattock, Ethan. 2022. "Free and Informed Elections? Disinformation and Democratic Elections Under Article 3 of Protocol 1 of the ECHR." *Human Rights Law Review* 22 (4). https://doi.org/10.1093/hrlr/ngac023.

Stiles, Kendall W. 2006. "The Dissemination of International Liberal Norms: The Case of the ECHR and the UK." *Canadian Journal of Political Science* 39 (1): 135–158. https://doi.org/10.1017/S0008423906500307.

Turton, Gemma. 2020. "Causation and risk in negligence and human rights law." *The Cambridge Law Journal* 79 (1): 148–176. https://doi.org/10.1017/S0008197319000898.

Waldron, Jeremy. 2020. "The rule of law." In *The Stanford Encyclopedia of Philosophy*, Summer 2020, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Weiß, Norman. 2017. "Origin and further development." Chap. 1 in *The Council of Europe: Its law and policies*, edited by Stefanie Schmahl and Marten Breuer, Part I: General aspects.

Winter, Christoph, Nicholas Hollman, and David Manheim. 2023. "Value alignment for advanced artificial judicial intelligence." *American Philosophical Quarterly* 60 (2): 187–203. https://doi.org/10.5406/21521123.60.2.06.

WJP (World Justice Project). 2022. *Rule of Law Index® 2022.* Published by the World Justice Project (WJP). https://worldjusticeproject.org/rule-of-law-index/global.

CHAPTER I

From *ought* to *is*
On autonomous weaponised reason

In this chapter, I introduce the *threat of algocracy* to *Enlightenment's legitimacy* paradigm in the context of the 5^{th} *Industrial Revolution* (INDUSTRY 5.0), why we should *resist* to that threat, and which *logical* requirements should legal ALGOAI models have in order to overcome it. More precisely, in §1, I introduce the concept of *evaluative judgements* which constitutes the core of ALGOAI engineers' practice as well as the main challenges ALGOAI engineers face when performing such judgements. It is an important section since it provide the terminology that will be used in the rest of the Thesis. In §2, I introduce the concept of *legitimacy* and how it is related to *Enlightenment*. I proceed by arguing that the value of *rule of law* entails legitimacy requirements for ALGOAI models which are more or less universal across different political orders. I continue by identifying more legitimacy requirements and the *priority relations* among them for the European and similar political orders. I further argue based on the foregoing legitimacy requirements how the separation of powers should be reconceptualised to *legitimately* facilitate the rise of the new unelected authority of ALGOAI engineers. In the last section of this chapter, §3, I introduce what is *algocracy*, what is its impact on the legitimacy of political orders, why & how it threatens Enlightenment's legitimacy paradigm in the context of INDUSTRY 5.0 and SOCIETY 5.0, i.e., the new social order that is predicated on the disruptions caused by INDUSTRY 5.0. Finally, I introduce *Ismene's dilemma*, the dilemma of whether we should accept or reject the post-Enlightenment legitimacy paradigm, and I argue why & how we should reject it.

I.1 On values

Common wisdom says that the proposition "*If the President issued a decree, then the President issued a decree.*" is true in virtue of its *logical form*: if X is true, then X is true. As logicians fancy to say, this proposition is true in every possible world; no matter how a world looks like, as long as the laws of classical logic hold, it is *objectively* true. Common wisdom also says that "*The President issued a decree.*" is true in virtue of the President actually having issued a decree. This proposition is true in all possible worlds where the *object* "*President*" behaved in the designated way; in those worlds, this proposition is *object*-ively true.[1] And then, common wisdom says that there are propositions like "*The President issuing a decree is undemocratic.*". Propositions that are *not* objectively true in *any* possible world.

What differentiates the last proposition in terms of objectivity is that it is a proposition about *judging* whether a *value* (the value of democracy) is *applicable* to a particular case (the President issuing a decree). Such a judgement is what is called in the literature an *evaluative judgement*.[2] For common wisdom, values are *non-objective* and ergo

[1] This position is many times rooted to the etymology of the word "***object*-ivty**": a proposition about an *object* is *object*ively true *if and only if* that object behaved in the way *described* by the proposition. And hence the intuition behind *objectively true* propositions is that they are *accurate descriptions* of the world (Putnam 2002, p.33, *cf.* Mulder, n.d.).

[2] *See e.g.* van Roojen 2018, §1.1; Stavropoulos 1996; Capaldi 1998, §9, footnote 11. It is also common to use "*value judgements*" (*see e.g.* Dworkin 2011; Putnam 2002; Alchourrón 2015; Vibert 2007, p.2; MacCormick 1992, p.182). However, as Sen 1967 remarks, value judgements constitute a superset of evaluative judgements. More precisely, apart from evaluative judgements, value judgements include purely *prescriptive* judgements. I.e., judgements about what *should be* the case according to a system of values like "*Capital punishment* ***should*** *be abolished.*". However, evaluative judgements like the examples of this paragraph have also *descriptive* content, if not *solely* descriptive content. We are concerned about what *is* the case (e.g., "*Capital punishment* ***is*** *barbarous.*"). Such descriptive judgments may also entail prescriptions. E.g., by saying that capital punishment *is barbarous*, one may also prescribe the abolition of capital punishment (*ibid.*, pp.46-47; *cf.* Putnam 2002, pp.67-70). Or by arguing that a robot judge *is fair*, one may also prescribe its use by national courts (e.g., Ulenaers 2020). By choosing the more precise "*evaluative judgements*", I want to emphasise that the ALGOAI engineers are concerned with *evaluating* whether an AI model abides by certain values and *not* which *should* be those values. The latter is a matter of politics (and ethics).

judgments about the applicability of values are *non-objective*. Examples of evaluative judgements are judgments about *moral values* (*moral judgements*) like simplistic bad-good dichotomies (e.g., "*Murder is* ***bad.***" and "*Praying is* ***good.***") or the more complex "*Not presuming innocence is* ***not just.***",[3] judgements about *epistemic values* (*epistemic judgements*) like "*The application of the law was not* ***foreseeable.***" and "*The defendant was found guilty beyond a* ***reasonable*** *doubt.*", judgements about *aesthetic* values (*aesthetic judgements*) like "*The simplicity of mathematics is* ***beautiful.***", and judgements about *political* values (*political judgements*) like "*The court did not preserve its* ***independence.***" or "*Presidential decrees are* ***undemocratic.***".

Evaluative judgements constitute the core of AlgoAI engineers' practice. Whenever AlgoAI engineers introduce an AlgoAI model (e.g., a robot judge), they essentially perform the following evaluative judgment:

The proposed AlgoAI model can exercise power ***legitimately***.

As we will see later on, evaluative judgements about the value of *legitimacy* are grounded on further evaluative judgements about the AlgoAI model's *realisation* of other values like the values of *democracy*, *rule of law*, *human rights*, *justice*, *transparency*, *accountability*, *reasonableness*, *forseeability*, etc. Engineering AI to abide by specific values is known in the AI literature as *value alignment* (Winter, Hollman, and Manheim 2023). When aligning AlgoAI models towards specific values, AlgoAI engineers are faced with the following *two-dimensional* challenge that I will call the ***objectivity challenge***: *if a value is not objective, how can we evaluate whether evaluative judgements of the said value hold?*[4] The *if*-clause constitutes the *ontic dimension* of the objectivity challenge: are values objective? The *then*-clause constitutes the *epistemic dimension* of the challenge: if values are not objective, how can we *know* whether judgments about those values hold? One could try to avoid the challenge by arguing that values are indeed objective. But life is not that easy.

value alignment & the objectivity challenge

I.1.1 On objectivity

> "*Debates on what makes something real are as old as time. It's something you'll have to decide for yourself, but I consider the matter settled.*"
>
> *Atlas in One Piece Chapter 1062*
> Eiichiro Oda, 2022

The discussion about evaluative judgements is a discussion traditionally concerning *meta-ethics*.[5] In the discourse of meta-ethics, values are construed as *concepts* that *guide* our actions (Dworkin 2011, p.1 and p.160) or more colloquially *principles* that *guide* our actions (Winter 2016, p.464).[6] We will see later that such *value-driven* actions include both *preforming* an act (henceforth *positive* act or simply *action*) and *omitting* the performance of an act (henceforth *negative* act or simply *omission*). Henceforth, whenever I use "*act*", I will be referring to both positive and negative acts unless specified otherwise. In meta-ethics, an evaluative judgement is essentially construed as a question of whether an *object* (e.g., a President issuing a decree)[7] is *subsumed* by a concept (e.g., democracy). In more formal terms, whether a *term* is subsumed by a specific *predicate*, or in more metaphysical terms, the ancient old problem of whether a *particular* is an instantiation of a *universal* (Alchourrón 2015, §1; MacCormick 1992, §II; *cf.* §II.4.1.2). Whenever an object is subsumed by a concept, I will call that object a *realiser* of that value or I will say that it *realises* that value. E.g., according to the Rule of Law Index® 2022, Cyprus scored better in the overall index score than Croatia (0.68/1 over 0.61/1) and hence Cyprus is a *more*

[3] *Contra* to popular belief, in (meta-)ethics, bad-good dichotomies are anything but simplistic (Schroeder 2021, §1).

[4] I say "*an evaluative judgement* ***holds***" instead of "*an evaluative judgement* ***is true***" since there is a debate in the literature, probably the most important debate regarding evaluative judgements (Navarro and Rodríguez 2014, p.51), of whether evaluative judgements can take truth values (*see* the source of this debate, the so-called *Jørgensen's dilemma*: Jørgensen, 1937/1938; *cf.* Navarro and Rodríguez 2014, pp.37-38 and §2.3). I.e., whether evaluative judgements can be what is called *truth-bearers* (MacBride 2022). Ergo, if one accepts that there are correct and incorrect evaluative judgements (i.e., "*Capital punishment is not barbarous.*" is an incorrect judgement, while "*Capital punishment is barbarous.*" is a correct judgement), they will have to either resort to alternative conceptions of the concept of *correctness* (alternative to identifying correctness as *truth*; *see e.g.* Kelsen's alternative to truth concept of *validity* (Kelsen 1991)) or they will have to provide an adequate response to the objections of evaluative judgements being truth-bearers. By using "*correctness*" instead of "*truth*" I remain neutral regarding this debate.

[5] Note that for Dworkin's, a seminal philosopher of law whose methodology of performing evaluative judgements I partially adopt (§III.3), there is no such thing as "*meta-ethics*". The "*meta*" part is illusionary; a meta-ethics discourse is *still* an *ethics* discourse. There is no second-order level. Even though I tend to agree with Dworkin's arguments (*see* Ehrenberg 2008 and Shafer-Landau 2010 for critical accounts of Dworkin's arguments; the last citation concerns Dworkin's last defense of the rejection of second-order ethics found in Dworkin 2011), since "*meta-ethics*" is the standard term used in academia for this type of discourse, I will abide by academic tradition for reasons of convenience but not conviction.

[6] Henceforth, I will be using "*value*" and "*principle*" interchangeably.

[7] Note that in the context of logic, by "*object*" one means the *object of discourse* and hence one does not have to commit oneself to a particular philosophical conception of objecthood. They can stay as metaphysically neutral as logic is.

adequate realiser of the concept *rule of law* than Croatia (WJP 2022, p.10).[8]

Note that the foregoing terminology (*concept*, *subsumption*, *realiser*, etc) is extended to any type of discourse in meta-semantics, not *per se* meta-ethical discourses. For instance, we can ask questions about whether a pamphlet or a fashion magazine are adequate realisers of the concept of *book* (Dworkin 2011, p.158). This allows for a uniform conceptual framework among ethics and the rest of the language without turning the discussion about values to a *suis generis* isolated discourse. Despite the common terminology, when two interlocutors talk about a specific concept (e.g., the concept of *justice*) in meta-semantics, they do not talk *per se* about the same concept. For instance, a *legal* conception of justice is not *per se* an *ethical* one. For ALGOAI engineers, it is important to distinguish between *legal* and *non*-legal concepts since as we will see later on (§2.6.4, ¶3 and §2.6.7, ¶1; *cf.* §2.1, ¶7), for an ALGOAI model to be legitimate, it has to be designed based on *legal* conceptions of values, with many times rejecting or ignoring their ethical conceptions. Legal concepts are essentially other types of concepts (political, epistemic, ethical, and so on), but *institutionalised* ones, institutionalised in specific *legal traditions* and *areas of law* of those legal traditions (*cf.* Stavropoulos 1996, p.47).[9, 10] For instance, regardless of whether there is a "*correct*" (what one would call an *objective*) concept of causation, we will see in §IV.1 that there is a specific construal of the concept of causation in the Anlgo-American criminal law which differs from the construal of causation in the human rights law of the ECtHR legal tradition. Or even inside the Anlgo-American legal tradition, the concept of causation differs from criminal law to contract law (Moore 2009, pp.513-514). This *ambiguity of meaning* is essentially what induces the objectivity challenge.

ambiguity of meaning

More precisely, when it comes to legal values in the legal meta-ethics discourse, a value not being objective is construed as being *subject*-dependent: for different subjects a value has different meanings and hence there is not a unique conception of that value (Schroeter, Schroeter, and Toh 2020; Dworkin 2011, pp.157-158). Even in the *same* trial, different judges have different conceptions of the same value. E.g., in the *Perincek v. Switzerland (2015)* case, the judges disagreed on whether human rights being *universal* entails that the severity of their violation is contingent on historical, geographical, and time proximity. They argued about whether denying a genocide was less harmful to the reputation of the genocide's victims if there is historical, geographical and/or time distance between the time and place where the genocide happen and the time and place where one denies it. Will denying the Holocaust be less harmful in 2050 than today? Is it more harmful if it was denied in Benin than in Poland? And how can one draw such distinctions?[11] If we extend the discussion outside of courtrooms as we should,[12] the problem of objectivity becomes even messier. Different conflicting political, ethical, religious, and social views, different personal experiences, different cultures, different socioeconomic positions, and so forth, shape different conceptions of what is *justice*, *democracy*, etc. Even the supposedly more "*neutral*" epistemic concepts like the concept of causation are contestable as we will see in CHAPTER IV.

Consequently, the objectivity challenge is precisified to the following question: *which conception* of legitimacy should the ALGOAI engineers follow to design AI models? For instance, should it be the Shari'a-influenced of the Middle-East area or the secular European one; two legitimacy paradigms with significant differences in determining the balance between the severity of the punishment and the severity of an offense (Araujo 2022, p.102). One can pose the question of the ontic dimension in different overlapping levels like at an (inter)national level, at a regional level, at a universal level, at a specific trial, at a specific court's tradition, in the practice of a particular judge throughout their career, in specific areas of law and/or legal traditions across courts of those traditions, in a societal level, in the level of a particular moral and/or political theory, in the level of a religious tradition, and so forth. Which brings about the question of *which* level is the *correct* one. Why should an ALGOAI model be legitimate according to European liberal tradition and not according to the East Asian Confucian tradition (*see e.g.* Fung and Etienne 2022, §2.2 and Chu 2016) or according to some unique universal conception of legitimacy? This is not a question of mere armchair academic interest. The desire to establish legitimate authorities according to a specific legitimacy paradigm was and still is the fuel for severe post-Enlightenment upheavals, from the French Revolution to the Arab Spring and the war in Ukraine (*cf.* §2.3, ¶8; §2.5, ¶12). And despite its importance, it does not receive the attention it needs, at least in the public

ontic dimension of the objectivity challenge

[8] The characterisation "*realiser*" is borrowed from Schroeter, Schroeter, and Toh 2020 whose method of performing evaluative judgements I partially adopt in CHAPTER III. On a similar note, the term "*adequate*" from "*adequate realiser*" is borrowed from the literature on *explication*, another method of acquiring knowledge about concepts that I use in CHAPTER III.

[9] In the context of this Thesis, it is sufficient to construe *legal tradition* as a tradition that includes "*deeply rooted, historically conditioned*" practices of *interpreting* and *applying* the law (Joutsen 2010, p.67). For instance, the *Anglo-American legal tradition* includes the practices of interpreting and applying the law in the UK, the Commonwealth, and the US (Moore 2019, §1). More details about what are the *interpretation* and the *application* of the law throughout the Thesis, especially in *fn.* 26, and §II.4.1.2 where I show how *logic* can explicate them.

[10] Examples of *areas of law* are contract, constitutional, criminal, tort, and human rights law.

[11] *See e.g.* the *dissenting opinion* of judges Spielmann, Casadecall, Berro, de Gaetano, Sicilianos, Silvis, and Kūris at pp.121-122 of *Perincek v. Switzerland (2015)*.

[12] I say "*should*" since according to the ECtHR's legal tradition, the content of human rights should always be updated *on par* with the current views of European societies, the so-called *European consensus* (Dzehtsiarou 2011; Dzehtsiarou 2015, §2.2; *cf.* §II.4.1.2).

discourse. As Floridi remarks, one of the biggest challenges of the current technological revolution is becoming aware of all those major changes in our newly emerged technology-infused life:

> "*As in a classic Renaissance house, we now inhabit the piano nobile, the upper, noble floor, not even knowing what happens in the ground floor below us, where technologies are humming in the service rooms.*"
>
> Floridi 2014, p.37

Apart from the ontic, there is also the *epistemic dimension* to the objectivity challenge. The objectivity challenge question is a *how-to* question of *evaluation*: "*...how can we evaluate whether evaluative judgements of [a specific] value hold?*" In other words, it is a question about *methodology*: *via* which *methods* ALGOAI engineers can determine whether an evaluative judgement is satisfied or not. This distinction between an ontic and an epistemic dimension is inspired by Benacerraf's position that an adequate answer regarding truth in mathematics should account for both an adequate theory of truth (*why* mathematical propositions are true) and an adequate epistemology (*how* do we acquire knowledge about those truths) (Benacerraf 1983). Benacerraf argued that philosophers of mathematics were able to satisfy one of the two dimensions, what became known in the literature as Benacerraf's dilemma (Hale and Wright 2002). This parallelism between meta-ethics and philosophy of mathematics is quite common in literature since both disciplines have as objects of inquiry *abstract entities* like numbers and values.[13] Benacerraf's dilemma sheds light on an important aspect regarding the objectivity challenge: one can not answer adequately one horn of the challenge without adequately answering the other. An adequate answer about the epistemic dimension should also be an adequate answer to the ontic dimension and *vice versa*. And *contra* to mathematics, in ALGOAI engineering, this requirement of answering both horns at once is not only of philosophical concerns, but it is of substantial *pragmatic* concerns. I doubt that Pelerrman concerned himself with questions about *what numbers are* when proving the Poincaré conjuncture. But I do not doubt that for the government of Estonia to design a robot judge that makes legitimate judgements (Niler 2019 *contra* Tuulik 2022), they should first answer *what legitimacy is* at least minimally.

epistemic dimension of the objectivity challenge

In the following section, I argue why an ALGOAI engineer can not avoid answering the ontic question of the objectivity challenge like one does in mathematics and in other non-philosophical disciplinary practices like law, medical science, and civil engineering. To do so, I need a generalised form of the objectivity challenge and its two dimensions that is applicable in all those disciplines and to which the objectivity challenge of evaluative judgements is a special case. That form is the following:

(GOC) Generalised objectivity challenge: *If a concept (e.g., tiger, justice, cancer) is not objective, how can we evaluate whether judgements of the said concept hold?*;

(OD) Ontic dimension: *Is there a correct conception of a concept?*;

(ED) Epistemic dimension: *Via which methods disciplinary experts can determine whether a judgment about a concept holds or not?*

I.1.1.1 Benacerraf's curse

Prima facie, it seems that in non-philosophical disciplinary practices like AI engineering, the practitioners do not have to concern themselves with ontology in order to evaluate the truthfulness of a judgement. Mathematics, physics, medical science, law, sociology, electrical and mechanical engineering, all of them have conflicting metaphysical theories about their ontology, and yet, such metaphysical disputes seem to have no influence in their practice. Courts convict criminals without concerning themselves about the ontology of justice, mathematicians continue to thrive even if metaphysical disputes about the truthfulness of $1+1=2$ are anything but settled, and quantum engineers are demolishing again and again the classical binary 0-1 computer paradigm despite all the metaphysical mess quantum physics brought on the table (they actually turned the table upside down). Even if the foregoing examples are accurate and indeed theories of truth do not impact disciplinary practices, in ALGOAI practice things are different.

In mathematical practice, a truthfulness of a proposition is decided in reference to a set of inference rules and a set of already accepted true propositions. As long as those rules are applied correctly to those propositions, the

[13] For instance, Stavropoulos draws a parallelism between the debate about objectivity in philosophy of law with the debate between realism and anti-realism in philosophy of mathematics (he refers particularly to Michael Dummett's anti-realist position of mathematical intuitionism). He does so in order to argue about both the global nature of the objectivity debate and its specificity to particular disciplines (law *v.* mathematics *v.* ethics, and so on) (Stavropoulos 1996, pp.5-6). The reader may also be interested in Leibowitz and Sinclair's 2016 "*Explanation in ethics and mathematics: Debunking and dispensability*" where the authors provide a defense of realism in both ethics and mathematics by using similar indispensability arguments from both disciplines.

conclusions they yield will always be true. Those rules and those propositions may differ in different areas of mathematical practice. E.g., one day the same mathematician will prove that a mathematical proposition is true in classical mathematics and the other day that it is false in intuitionistic mathematics. Metaphysically, it may be that only one of the two cases holds. But the mathematician does not care about ontology. Mathematical practice is a *formalistic* practice in which the mathematician takes sets of rules and sets of propositions as *given* without questioning their metaphysical status.[14] That can not be the case for ALGOAI engineering. It can not be the case that *both* police drones that are designed based on the value of racial superiority and police drones that are designed based on the value of racial equality are legitimate. Only one of the two conflicting evaluative judgements is true. And the same goes for the inference rules used to derive evaluative judgements; the judge of the Anglo-American criminal law tradition can not use the causal inference rules of the ECtHR to establish causation and responsibility between the actions of the defendant and the harm of the victim (*see* §1.1, ¶2).

It seems that ALGOAI engineers can not ignore ontology in the way that mathematics does. They can not take certain propositions and inference rules for granted. How about avoiding ontology in the way that the rest of the disciplinary practices used as examples (law, empirical sciences, engineering, and so on)? In those disciplinary practices, there is generally *no* set of inference rules that always yields true conclusions. There are though methodologies that can be used to infer true conclusions *most* of the time. E.g., in a trial, DNA tests are used to judge someone guilty *beyond reasonable doubt*. "*Beyond reasonable doubt*" entails that there is always the possibility of the court being wrong and the defendant being innocent. It can be that DNA similarities are coincidental or that the DNA evidence is planted or that DNA samples were swapped or that the DNA evidence was self-materialised out of thin air. But whenever all precaution measures and safeguards have been followed and whenever all other evidence are taken into consideration then the probabilities of those possibilities happening are so low that they are deemed *unreasonable* (*see e.g.* Meester and Slooten 2021, §§10-11, and pp.224-225). Similarly, biopsies or pregnancy tests have a certain percentage of misdiagnosis (*see e.g.* Dirks et al. 2023 and Anderson and Ghaffarian 2023; DeLaney and Wood 2021 respectively). Or the estimation of an engineer that a building can withstand a 7-Richter magnitude earthquake is just that: an estimation, not a certainty. In other words, whenever disciplinary experts want to know the truthfulness of a proposition, the normative is to follow the methodology that is the most probable to yield true conclusions.[15]

To compare methodologies and see which one is more probable to yield true conclusions, we need to already have some *prior minimum non-subjective* knowledge about the involved ontology (following Dworkin 2011 (pp.160-163) I will call such knowledge *paradigmatic* knowledge). I.e., we need to be able to *know* certain propositions about that ontology so as to be able to evaluate how often and under which circumstances the conclusions of a proposed methodology are true. And since we want this evaluation to be non-subjective, there needs to be a strong consensus about the truth of those propositions. We can verify the accuracy of a pregnancy test by comparing it with another test that is already considered to be accurate enough (the so-called *criterion standard*) (Fromm et al. 2012; Anderson and Ghaffarian 2023). Or we can evaluate whether buildings can withstand earthquakes by constructing full-scale replicas and testing them on shake tables that reproduce seismic events of different magnitudes (Miglietta et al. 2021). Even in physics where it is common practice to postulate the existence of *unobservable* entities, in order to evaluate whether a methodology produces true propositions about the unobservables, we exploit our knowledge about the observable aspects of the physical world "*without interacting with the [unobservable] object in question*" (Arvidsson-Shukur, Gottfries, and Barnes 2017, p.1). In order to perform all those evaluations, it is *epistemically necessary* for the experts performing the evaluation to have firstly conceded on some paradigmatic knowledge like the propositions produced by the criterion standard in medical science, the operation of seismic tables in civil engineering, or the knowledge about observables in physics.

Summing up, we can not use a methodology to acquire new knowledge (i.e., answer the *epistemic* dimension of the objectivity challenge) unless we can test that methodology with previously acquired knowledge (i.e., provide a minimal answer to the *ontic* dimension of the objectivity challenge). In mathematical practice, we can do so because we do not care about metaphysics. But in the rest of the disciplinary practices mentioned so far, it turns out that we *do* care about ontology, at least minimally. We care about *what* is the case in the *actual* world. We care whether one is pregnant, whether they have cancer, whether they killed someone, whether the building can withstand strong earthquakes, or whether a quantum computer can work. And similarly, we *care* about whether a robot judge can be bribed, whether a police drone considers a civilian suspicious based on

Benacerraf's curse

[14] Apart from testifying about the formalistic character of mathematics based on my personal experience as a mathematician, the reader can have a look at journal of mathematics to ascertain the veracity of my argument.

[15] I am saying "*the normative*" because there can be factors that do not allow to choose the optimal option. E.g., limited resources or ethical restrictions like restrictions on human and animal experimentation. For instance, knowledge about the causal interaction among ensembles of neurons in the human brain can not be acquired by directly intervening in such neurons. Neither we can do so to animals with similar-to-human brain structure like male rhesus monkeys. Consequently, we will have to resolve to other alternatives like *observing*, and not directly *intervening*, the neural activity of humans or male rhesus monkeys (Chen 2021, pp.522-524).

their race, and whether spyware collects personal data without our consent. I will call the epistemic necessity of paradigmatic knowledge *Benacerraf's curse*.

The practical importance of the necessity for prior paradigmatic knowledge can be seen in both the practice of engineering AI & the practice of judicial decision-making. Regarding AI engineering, take the example of what is probably the most funded and the oldest AI real-life application, AI that performs medical diagnoses (computer-aided diagnoses (CADx)). Contemporary state-of-the-art CADxs are trained to identify patterns in past diagnoses that they later use to diagnose new cases (*see* e.g. McKinney et al. 2020, one of the most famous CADx cases that outperformed human express in breast cancer detection (Sample 2020; Lemke 2020; Goodwin 2020)). This pattern identification is possible *only if* there is already a consensus among the experts about the diagnoses used to train the AI since those paradigmatic diagnoses are what constitutes the criterion standard for evaluating the AI's accuracy. An example of a problematic case of designing such medical AI due to lack of paradigmatic knowledge is the CADx of polyps (Mori et al. 2022, p.370). Regarding judicial decision-making, human rights constitute the paradigmatic prior knowledge that should be the foundations of any interpretations and applications of the law; they are the *bear minimum* that all judgements should abide by (§2.6). Consequently, disagreements about the content of human rights can have a substantial impact to on the interpretation & application of the law. One can always ask the *regressive* question of which is the prior paradigmatic knowledge that justifies propositions about human rights. One answer is that human rights are traditionally construed as self-evident truths as famously stated in US Declaration of Independence (4 July 1776): "*We hold these Truths to be **self-evident**, that all Men are created equal, that they are endowed by their Creator with certain **unalienable Rights**, that among these are Life, Liberty, and the pursuit of happiness*".[16] Having said that, one can always question what makes those rights self-evident. Why is the right to freedom of thought or the right to life self-evident? Interestingly, one can once more draw parallels between meta-ethics and the philosophy of mathematics by comparing the self-evidency of human rights with the self-evidency of mathematical axioms (Shapiro 2009, pp.204-205).

In any case, whatever one's position is on the self-evidency of human rights, the conclusion is that what differentiates evaluative judgements from the examples of the non-evaluative judgments like "$1 + 1 = 2$." and "*Melina has cancer.*" is that the truth of the latter is generally grounded on paradigmatic knowledge countering Benacerraf's curse. Since the amount of expert consensus is traditionally higher in the cases of non-evaluative judgment like "$1 + 1 = 2$" or "*Melina has cancer.*", one could justifiably propose to reduce the question of whether an evaluative judgement is true to a question of whether such non-evaluative judgements are true. To be *on par* with the literature on philosophy of science, I construe the latter as *factual judgements* (or *factual propositions*). For instance, the proposition "*If the President issued a decree, then the President issued a decree.*" and "*The President issued a decree.*" (§1.1, ¶1), mathematical propositions or logical propositions ("$1 + 1 = 2$", "$p \vee q \rightarrow p$", or propositions from empirical sciences ("*Melina has cancer.*") are *factual* propositions. Factual judgements are juxtaposed to evaluative judgements constituting the so-called *fact/value dichotomy* (Putnam 2002; §2.3, ¶7; *cf.* §2.5, ¶4).

I.1.2 "*Factualising*" values

Analytic philosophy, political and social sciences, and especially economics have provided many such attempts to reduce evaluative judgements to factual judgements (Putnam 2002). The WJP's rule of law index or EIU's democracy index are such attempts. For instance, the evaluative question of whether a state abides by the rule of law is reduced to factual questions about whether the general public can request and receive information from state authorities, whether they need to bribe to get it, or whether the waiting time for receiving the information is lengthy. More precisely, the WJP assigns *numerical scores* to states depending on how adequate they realise the rule of law. The final score is calculated by aggragating the numerical scores collected by Qualified Respondents' Questionnaires (QRQs) that were given to experts (3.600 experts from 140 coutnries which is about 26 experts per country) and by General Population Polls (GPPs) that were given to each state's general public (the default being 1.000 respondents per country and 340 questions). In total, 500 variables are used in calculating the numerical scores. Those numerical scores are further processed using standard statistical methods (e.g., normalising the scores using the min-max method, testing annual differences using t-tests, deleting outliers using the z-score method). The WJP further cross-checks the processed scores with 70 third-party sources to identify biases and errors and they conduct sensitivity analysis with the aid of the Econometrics and Applied Statistics Unit of the European Commission's Joint Research Centre.[17] In other words, *prima facie*, the evaluative judgement of

[16] Emphasis added. As I argue in §2.3, ¶5 and §2.7, ¶1, Enlightenment's *natural rights* like the Declaration's rights to "*[l]ife, [l]iberty and the pursuit of happiness*" are precursors of the post-WWII concept of human rights which can also be construed as self-evident (*see e.g.* Etzioni 2010; Shapiro 2009, p.205).

[17] All details about the methodology can be found in WJP 2022, pp.182-185. The QRQ and GPP questionnaires can be found in

whether a state has realised adequately the value of rule of law is reduced to *factual propositions* and to *standard methodologies* of processing such factual propositions.

However, even if one accepts that such a reduction is valid, this reduction is possible *only if* one makes certain *background* evaluative judgements. If one adopts the position that the truth value of an evaluative judgment $J(v)$ about a value v depends on certain factual judgements f_i, one still makes a judgement about v; arguing that v is reducible to f_i is *still* a judgment about v. Arguing that bribery of public authorities undermines the rule of law is still a judgement about the the rule of law. Similarly about the choice of methodologies: arguing that t-tests are an appropriate tool to evaluate whether there are annual changes in the realisation of the rule of law is still an evaluative judgement about the rule of law.

This is no new argument. Putnam 2002 (pp.55-56) provides the example of the *Pareto optimality*, a quite popular *factual* measure of the value of *optimality* in economics and other disciplines. Assume a set of actors (e.g., humans) and a set of resources (e.g., income). The Pareto optimal allocation of the resources to the actors is the one which satisfies the following construal of the concept of *optimality*: one can not make a change to the allocation of the resources that will move an actor to a better position without moving another actor to a worse position. Using Pareto optimality as a measure of optimality presupposes *inter alia* the background judgement that every actor has an *equal right* to maximize their position which is a textbook evaluative judgement. And of course, by being an evaluative judgement, it is by default vulnerable to controversy: "*Defeating Nazi Germany in 1945 could not be called Pareto optimal, for example, because at least one agent-Adolf Hitler-was moved to a lower utility surface*" (*ibid.*, Putnam 2002, p.56).

Even though attempts to reduce evaluative judgements to a set of judgements that are *exclusively* factual are futile, identifying a *checklist* of factual judgements that influence the realisation of a value (e.g., whether the rule of law index is higher than 0.7/1 or whether the p-value of a t-test is < 0.01) is both *epistemically* and *pragmatically* useful. Such a checklist is what is called an *operational definition* of that value, where the intuition behind operational definitions is that "*we do not know the meaning of a concept unless we have a method of measurement for it*" (Chang 2021). Reducing the definition of a value to an operational definition is *epistemically* useful since the more consensus experts have on the knowledge they produce the more *new* inter-subjective knowledge they can acquire as exhibited by Benacerraf's curse. It is also *pragmatically* useful because it allows us to identify the similarities and differences between different paradigms of a value (e.g., different paradigms the value of *legitimacy*) and hence to find common ground among those paradigms. That common ground can be used to *develop* and *commercialise* AlgoAI models across regions with different paradigms (more on §2.6.2). Finding common ground *via* an operational definition was exactly what the Venice Commission[18] did in 2011, when they conferred to decide on a definition for the *rule of law* that can be used across the CoE members states. They concluded that the only way to provide a rule of law definition compatible with the different European rule of law paradigms (e.g., UK's *Rule of Law*, Spain's *Estado de Direito*, Russia's *Правовое Государство* & *Верховенство Закона*, Germany's *Rechtsstaat*, and France's *État de droit* (CDL-AD(2011)003rev, pp.3-5)) is to provide an *operational definition*.[19] 5 years later, the Venice Commission provided an even more detailed checklist (CDL-AD(2016)007; *see also* footnote 19). Having said that, their checklists do not consist exclusively of measurable or factual requirements, but of evaluative requirements as well. This is not an issue though. We can construe the operational content of a definition as a spectrum: the further towards the operational non-evaluative side a definition, the more likely to be an inter-subjective consensus about the knowledge acquired by that definition.

the perks of *operational* definitions

Summing up, it is advisable for AlgoAI engineers to provide operational definitions (*checklist*) for the different values AlgoAI models need to be aligned towards. The requirements of those definitions, the checkboxes, will be called *operational requirements*. The adequacy with which a model realises a value will be based on the degree that it satisfies the operational requirements. We will see that such requirements can be about both positive and negative acts. I will call them respectively *positive* and *negative requirements*. For instance, the *High Contracting Parties* (*HCPs*) have the positive obligation to take specific measures that protect human rights (e.g., introducing appropriate legislation) and the negative obligation to abstain from interfering with such rights (e.g., abstaining from censoring free speech) (Lavrysen 2016).

operational requirements

https://worldjusticeproject.org/2021-wjp-rule-law-index-questionnaires (accessed 10 February, 2023).

[18]The *Venice Commission* (*European Commission for Democracy through Law*) is one of the CoE's organs that is responsible for the realisation for the rule of law (*see* the classification of CoE's organs in the tabs of their website (accessed on 10 February, 2023): https://www.coe.int/en/web/portal). One may wonder why the Commission is classified under the "*rule of law*" and not the "*democracy*" tab despite having "*democracy*" in its full name. This should not come as a surprise since as we will see in §2.4, ¶2, for the European order, the realisations of three legitimacy pillars are conceptually interdependent.

[19]In ¶70 of CDL-AD(2011)003rev, the Commission explicitly acknowledges the rule of law requirements it provides as a *definition*. In ¶4 of its website (https://www.venice.coe.int/WebForms/pages/?p=02_Rule_of_law&lang=EN), they characterise it as "*operational*". In ¶5 of the same website, they give the same characterisation ("*operational*") to the more detailed checklist they drafted 5 years later (accessed on 10 February, 2023).

In what follows, I will provide operational requirements for the value of *legtimacy* that should be followed by AlgoAI engineers. But first things first, I need to explain what *is* legitimacy.

I.2 On legitimacy

I.2.1 On order

> "*Men may, of course, have order without liberty, but they cannot have liberty without order.*"
>
> *Political order in changing societies, pp.7-8*
> Samuel P. Huntington, 1968

Huntington argues that there can not be liberty without *order*, referring to the *political* concept of order. Similarly, there can not be rights, democracy, rule of law, and let alone *legitimacy* without a pre-existing political order. It is *inter alia* a matter of conceptual priority; we first form a political order and then we ask whether that order is legitimate, whether it abides by the rule of law, whether it is democratic, and whether it protects human rights (*cf.* [§2.3, ¶]). Thus, before arguing what is the quality of legitimacy that an order *can have*, we need to start with the conceptually prior question of what an order *is*.

What is order?

Order can be construed in two dimensions: (α) *structural dimension*: a set of *actors* and *relations* among those actors that constitute a certain *arrangement* (or *structure*).; (β) *functional dimension*: the *ends* that the foregoing arrangement is expected to realise. The arrangement is arranged in the way it is arranged with the intention to realise those ends (Devetak and Dunne 2005, p.613). Following the Weberian terminology,[20] I will say that order is *oriented* towards its ends (Weber 1947, pp.124, *see also* pp.91,115). E.g., the ECtHR and the 46 HCPs are actors that form a specific arrangement. The relations among those actors that structure that arrangement are relations of *authority*: the ECtHR has *judicial authority* over the HCPs. As argued in the Introduction, the functional dimension of that arrangement is *inter alia* the protection of human rights, democracy, and the rule of law in the territory[21] of the HCPs. This order is *oriented* towards realising those ends

What are actors?

We saw that an order consists of *actors* and *relations* among those actors. This brings up the questions of what exactly is an actor and which are the relations among them. Let's begin with the first question. The *actors* of an order are called "***act-ors***" because they have the capacity to perform *social acts* (and hence one can also call them *social actors* that are related *via social relations* so as to form *social orders*). Social actions are actions that have two properties: (a) the actors attach a specific *meaning* to those acts.; (b) to perform a social act, an actor takes into account past, present and future actions of other actors (Weber 1947, pp.88,112). Applying the law, buying birthday presents, and ignoring a WhatsApp text are all social acts. Note that social acts can be both positive and negative (*ibid.*, p.112). Depending on the discipline and the practical purposes of the discourse, an actor can be construed only as a *single* individual (that should usually be the case for social science according to Weber (1947, pp.101-102)) or as a collection of individuals (e.g., a state, a court, an army, a family, an NGO). Construing collections of individuals as actors is quite common in law and politics (*ibid.*). E.g., the ECtHR (a collection of individuals) decides whether a state (another collection of individuals) violated a human right protected by Convention (Articles 33 & 34) or in international relations experts and diplomats are concerned with the interaction among states.

What are state-actors?

To avoid terminology overloading, I will label all actors that act on behalf of the state (courts, ministers, public servants at administrative offices) as *state-actors*. No longer confusing verbosity like "*state organs*", "*state institutions*", "*state organisation*", "*state representative*" and so forth. Everything will be a state-actor unless there is the need to specify otherwise (there is no such need). Note that if a state-actor does not act on behalf of the state, their acts are not considered *state-acts*, and hence, in this context, they are no longer considered state-actors. For instance, a judge may buy a birthday present for their daughter or vote in a municipal election. They do so as a *non*-state-actor. But when they issue a verdict or adjudicate a divorce, they act as a state-actor. When a state has violated the Convention, it is a state-actor that did the violation, from the medical personnel of a public hospital (*see e.g. Lopes de Sousa Fernandes v. Portugal, 2017*, ¶147) to the state's law enforcement and regional courts (*see e.g. Beizaras and Levickas v. Lithuania, 2020*, ¶155) (*cf.* Stoyanova 2018, pp.318-319 and §7.C). Note that I will further abuse the term "*state-actor*" to notate actors that are formed by mutual agreement of state-actors (e.g., the CoE, the ECtHR, the UN). I will do so by using a qualitative adjective before "*state-actor*" (e.g., the UN is an *international* state-actor).

[20] Max Weber (1864–1920) is considered the founding father of contemporary social science and a point of reference regarding the concept of legitimacy (Beetham 2013, p.8; Peter 2017, §1; S. H. Kim 2022, §1).

[21] Regarding territorial conflicts of the ECtHR's jurisdiction *see* ECtHR's Registry 2022a and Press Service of the ECtHR 2018.

Before explaining which are the relations among actors that form orders, I would like to draw a distinction between *actors* and *agents*, a distinction that can prevent us from deriving misleading conclusions with adverse consequences for AlgoAI. A key challenge in interdisciplinary practices is the conflation of terms: multiple disciplines use the same term but in different ways. I.e., another ambiguity of reference problem that is addressed in chapters ii & iii. In our case, the ambiguity of reference concerns the term "*agent*". For social, political, legal sciences,[22] and the like, one uses the term "*agent*" instead of "*actor*" for a subset of actors that have certain extra properties like *self-reflection* and *sense-of-self* (Voronov and Weber 2020, p.874; *cf.* Schlosser 2019). In those disciplines, those extra properties *do* matter.[23] In logic and sometimes in AI modelling, whenever we want to model interactions among any type of "*individuals*" that have the capacity of acting (any type of *act*, not *per se* social acts), we use the term "*agent*". It is an umbrella term without too much philosophical consideration behind it. It is similar to the use of the term "*object*" (*fn.* 7). *Contra* the discipline of logic, in this Thesis, I will reserve "*agent*" only for the cases where I want to highlight certain properties of an actor's agenthood. For instance, in §3.2.1.1, ¶7, I argue that AlgoAI is not an *epistemic agent*. This is of prime importance since AlgoAI having epistemic agency abilities entails obligations for the AlgoAI like refraining from using its epistemic abilities to do harm. Consequently, AlgoAI as an epistemic agent can be burdened with *legal* and *moral* liability which would have burdened its engineers and its users had the AlgoAI not been an epistemic agent (*cf.* Russo 2022, §9.5.2). Cases like Sophia the robot to which Saudi Arabia granted the *right* of citizenship in 2017 (Walsh 2017) are nothing more than crude efforts to make headlines. The fact that contemporary AI is not an epistemic agent was the reason why I used the term "*subject-dependence*" in §1.1, ¶3 instead of "*mind-dependence*" as many philosophers of (meta-)ethics do (*see e.g.* Stavropoulos 1996, §4.2). *Mind*-dependency would entail that contemporary AI *has* or *is* a mind. And that is in no way I claim I wanna make. Not only because it is factually untrue, but also because stretching the definition of "*mind*" would then leave open the doors for AlgoAI having epistemic agency (more on §3.2.1.1, ¶7).

agents *v.* actors

Now it is time to explain which are the *relations* that make up an order. In the general case, an order is made up by *social relations* among social actors, where a social relation "*consists entirely and exclusively in the existence of a probability*" that a social actor will socially act by considering actions of other actors (Weber 1947, p.118). E.g., the ECtHR and the HCPs are socially related because there is a possibility that the ECtHR will apply the law (social action) by considering the actions of the HPCs. In the context of this Thesis though, we care about a particular type of social relations, *political relations*. More precisely, when a social action is a *political act* (e.g., applying the law is a political act but buying a birthday present for your best friend is not), the relation grounded in the possibility of that act happening is a *political* relation and the actors partaking in this relation are *political* actors. I construe as *political order* the order that consists of political actors and political relations among those actors. Henceforth, unless specified otherwise, by "*order*" I will mean *political order*. What exactly constitutes a *social* and what a *political* act is not of relevance for the Thesis. After all, precisifying this distinction would require a Thesis on its own. The political acts that are of relevance for this Thesis are those of *exercising political power* like judicial, legislative, or executive power (more on §2.2, ¶4).

social relations *v.* political relations

Before concluding, I will provide a *typology of orders* that will be of use, especially for the definition of the value of *legitimacy*. A political order that is ordered in a way that adequately realises the ends of its functional dimension is a *well*-ordered political order (Devetak and Dunne 2005, pp.613-614, Brownsword 2022, p.33) *contra* a *dis*ordered political order that fails to do so (Barma 2016, p.45). Note that some (e.g., Huntington 2006; F. Fukuyama 2011, 2014) use "*order*" as a synonym to "*well-ordered order*" (Barma 2016, p.45). It is important for this Thesis to maintain the distinction between the two in order to argue when AlgoAI makes an algocratic order more well-ordered and when it makes it more disordered. I also need to make two further distinctions of (well-)orderness. The first distinction is between what I construe to be a *descriptive* and what a *normative* definition of well-orderness. The definition I have laid out so far is the *descriptive* one: an order is a *good* order if and only if the ends of its functional dimensions, whichever those ends may be, are realised adequately. Those ends though may not be *good* ends. I.e., they may not be what is considered to be the *normative*. Considering this, I construe as a *normatively* well-ordered order the order that is well-ordered towards *good* ends. E.g., Brownsword distinguishes between a normatively well-ordered order (what he calls a "*just*" order) and a descriptively well-ordered order (what he calls a "*good order*") by distinguishing between obeying the law when it is *moral* and obeying the law for reasons of *legality* regardless of whether the law is moral or not just like what Socrates did when he drank the conium (Brownsword 2022, p.33-4; more about the value of *legality* in §2.4, ¶1). Finally, the second distinction I would like to make is that between a *politically ordered*

typologies of *order*

[22] To the infuriated philosopher of science, I kindly ask to be patient until chapter ii, §4.1.2, where I define what is "*legal science*".

[23] For instance, according to CoE's MSI-AUT, contemporary algorithmic decision-making systems (ADMs) *object*-ify the individual by stripping their *agent*-hood properties like morality so as to be "*...sorted, sifted, scored and evaluated [...] in ways that appear starkly at odds with the basic right of all individuals to be treated with dignity and respect, and which lies at the foundation of all* ***human rights*** *and* ***fundamental freedoms.***" (MSI-AUT 2019, p.36, emphasis added). I.e., the distinction between actors *with* agenthood and actors *without* agenthood is of prime importance when dealing with *human rights* concerns prompted by new technologies.

social order (henceforth simply *order*) and a social order that is *not* politically ordered (henceforth an *unordered* order). E.g., cyberspace, especially in its earliest days, is such an unordered order (Floridi 2014, pp.184-185; *cf.* §2.6.1, ¶¶5-7). There were no laws to stop the dissemination of revenge porn or to tax cryptocurrency transactions. The traditional political orders had to digitalise their political activity in order to exercise such powers (*see e.g. Volodina v. Russia (No.2), 2021* and EC 2023 respectively).

Having introduced the concept of *order*, I can now introduce the concept of *legitimacy*, a quality that characterises the legitimate political orders. Before doing so, I need to introduce two more concepts: *authority* and *power*.

I.2.2 On authority, power, and legitimacy

Let's have a look again at the example of the political order formed between ECtHR and the HCPs. As I argued in ¶2 of §2.1, they stand in an *authority relation*: the ECtHR has *judicial authority* over the HCPs. I.e., authority is an *asymmetric social relation* between two *actors* and we say that one actor *has authority over* the other actor (Liese et al. 2021, p.356). The latter actor is usually called *subject* and the former is usually called a *authority* as well (*see e.g.* Christiano 2020, §1). Hence, we should be careful not to conflate the authority-actor with the authority-relation. The ECtHR is an authority-actor that stands in an authority-relation with its subjects (the HCPs). If the social relation of authority is a *political* relation like the relation predicated on the possibility of exercising judicial power, we have a case of *political authority*. Henceforth, by "*authority*", "*power*", "*relation*", "*order*", I will mean their *political* counterparts unless stated otherwise explicitly.

What does it mean though to have *authority* over a subject? Let's take the example of a *state*. A state has authority over its subjects (citizens, corporations, associations, etc) whenever: (α) the state maintains *public order* (*ordre public*), where a standard definition of public order is the conjunction of public security (or public safety) (sécurité publique) and tranquility (tranquillité publique) (Gaudemet 2015, p.3). Operational definitions of *public tranquility* include the absence of widespread criminal, political, or other types of turmoils, while operational definitions of *public security* include the adoption of effective measures that prevent & combat such turmoil like successfully implementing processes of peaceful transfer of power or anti-terrorist operations (*cf.* WJP 2022, pp.18; MSI-AUT 2019, p.29).; (β) the state issues *rules* that impose *obligations* that the subject generally *obeys*. Conditions (α) and (β) constitute a *de facto* (i.e., a *descriptive*) account of authority similar to the *de facto* account of authority proposed by Enlightenment thinker Thomas Hobbes (1651) and the seminal philosopher of law John Austin (1832) (Christiano 2020, §1). The motivation for choosing this account of authority will be given at the end of this subsection (§2.2, ¶11).

conditions for *authority*

Now, the ECtHR is not a state, but it still satisfies the two conditions: (α) the ECtHR *does* protect the public order of the HCPs. For instance, multiple times in its case-law,[24] the Court has emphasised the Convention's role as a "*constitutional instrument of European public order*" (*see e.g. Loizidou v. Turkey*, ¶¶75,95; *Bosphorus Hava Yolları Turizm ve Ticaret Anonim Sirketi v. Ireland*, ¶15; *N.D. and N.T. v. Spain*, ¶110). Public order is also protected explicitly by the Convention (*see e.g.* ARTICLE 6 (RIGHT TO A FAIR TRIAL), ¶1 and ARTICLE 9 (FREEDOM OF THOUGHT, CONSCIENCE AND RELIGION), ¶2).; (β) the ECtHR *does* impose obligations to the HCPs *via* its judgements and the HCPs generally obey them.

The concept of *power* concerns condition (β). More precisely, there is anything but consensus as to what the concept of power is (*see e.g.* Allen 2022, §1; Brown, McLean, and McMillan 2018, pp.989-992). In its wider sense, power can be construed as the ability of an actor to realise their desires (Beetham 2013, p.43). In our case, a political authority desires to make its subjects to generally obey the rules it issues even when the subject does not want to obey them (Christiano 2020, §1). When an authority does so, we say that it *exercises (political) power* (*see e.g. ibid.*, §1.1,§1.3,§5.1). Traditionally, political power is construed in the three dimensions of Montesquieu's *separation of power* from his Enlightenment magnum opus "*De l'Esprit des lois*" (1748) (Sedley 2015, p.172 *cf.* Law 2022, p.1475): (α) *legislative power*: the political actor *makes* legal provisions;[25] (β) *judicial power*: the political actor *interprets* and *applies* legal provisions;[26] (γ) *executive power:* the political actor *enforces* the interpreted and applied legal provisions.

What is *political* power?

[24] *Case-law* refers to past court judgments used to make similar decisions in similar cases (Law 2022) (*cf. principle of equality* in §2.4, ¶2). This type of reasoning is usually called *analogical reasoning* (Alexander and Sherwin 2008, §III.I; Lamond 2016; *cf.* Walton 2002, pp.35-39) or *case-based reasoning* (*CBR*) (Bongiovanni et al. 2018, pp.53-57). *Contra* to case-law, *statute law* is the law found in the legislation introduced by legislative authorities (Law 2022).

[25] I construe "*legal provision*" as an umbrella term that refers to any type of *authoritative* text (e.g., laws, treaties, international human rights instruments) whose authoritative content is about regulating the behaviour of a group of agents. That group of agents constitutes the *jurisdiction* of the legal provision. I base this construal on Governatori, Rotolo, and Sartor 2021, p.664.

[26] I construe as *interpretation* of the law the determination of law's *content* (e.g., the ECtHR determining the content of the human rights' value of *universality* (*cf.* §1.1, ¶3; §2.7, ¶3)). As *application* of the law, I construe the decision of what the interpreted law dictates in particular cases. E.g., in the *Perincek v. Switzerland (GC, 2015)* case, the ECtHR judged that the interpretation of human rights' *universality* in conjunction with other interpreted values dictates that Switzerland violated the Convention by interfering with the applicant's right to

The purpose of the separation is to put in place a system of *checks and balances* which aims at restricting the abuse of power by political authorities (Beatson 2021, p.5). Apart from *ab*use of power though, checks and balances also prevent the *mis*use of power. I construe both abuse and misuse of power as cases where an authority exercises power in a way that goes against a political order's functional dimension. As *abuse* of power, I construe the exercise of power that undermines the functional dimension due to *ethical* reasons (e.g., a judge is being bribed). As *misuse* of power, I construe the exercise of power that undermines the functional dimension due to lack of *competency* (e.g., when a government fails to propose legislation that adequately protects human rights).

abuse *v.* misuse of political power

Summing up, I construe as *political authority* the asymmetric social relation between an authority-actor and a subject. That relation is predicated on the *possibility* that the authority-actor will perform a political *act* of exercising *power* towards the subject in conjunction with maintaining *public order*. To generalise this definition to any type of *social* authority, not *per se* political, I remove the condition of maintaining public order: a social authority relation is predicated *solely* on the possibility that the authority-actor will perform an act of exercising power towards the subject, i.e., solely on *making the subject generally obey rules even if the subject does not wish to do so*. Such social authority is the *expert* or *epistemic* authority (Liese et al. 2021, p.356). E.g., the authority of the medical expert that issues rules to their patients that they generally follow even if they do not want to do so. As we will see in §2.5, whenever the exercise of power by the medical experts contributes to the security of public order like in the case of the pandemic, then that social authority becomes *political*. Such epistemic political authority is the authority that AlgoAI and its engineers have (more on §2.5).

simply power

Now, I can finally introduce the concept of *legitimacy*. *Legitimacy* is a *mode* of exercising political power. I.e., a way of making the subject obey a rule even if they do not wish to do so (Hurd 2005, p.501; *cf.* Brown, McLean, and McMillan 2018, p.992). Typical means of exercising political power are *coercion* like the threat of imprisonment, *consent* like the parliaments of the HCPs voting in favour of ratifying the Convention, and *manipulation* like restricting freedom of the press (*ibid.*, pp.992-993; Hurd 2005, p.501). Legitimacy is not when the subject obeys a rule because they have been coerced or manipulated or consented to do so. A *legitimate* exercise of power is when a subject obeys a rule even if they do not wish to do because they *believe* (Weber uses "*vorstellung*" (1947, p.124)) that they ought to obey a rule (*ibid.*, p.501; *cf.* Brown, McLean, and McMillan 2018). Hence why Weber argues that legitimacy is equivalent to "*Legitimitätsglaube*" (*trans:* "*a belief in legitimacy*") (Beetham 2013, p.8; Peter 2017, §1). A subject may obey a rule out of coercion that they will be imprisoned but that does not entail that they *believe* that this rule ought to be obeyed. They deem this rule *il*-legitimate leading many times to the disruption of public order with massive protests, revolutions, and so on (*see* §2.3, ¶8; §2.5, ¶12). Or the parliament may have *consented* to ratify the Convention, but that does not entail that future parliaments or the people that voted the members of the parliament (MPs) *believe* that the Convention needs to be obeyed.[27] It can also very well be the case that the MPs that voted to adopt the Convention they themselves do not believe that the Convention ought to be obeyed; they ratified it for diplomatic or other reasons.

legitimacy as a mode of exercising power

We saw in §2.1, ¶2, that an order is ordered the way it is ordered in order to (I apologise, but I could not resist) satisfy the ends of its functional dimension. Consequently, the authority relations and the exercise of power that establishes those relations exist to satisfy those ends. E.g., a judgement delivered by the ECtHR is expected to realise the three legitimacy pillars. Considering this, people's belief that they should obey the rules of that order entails an acceptance of its ends. It also entails the belief that those rules indeed contribute to the realisation of those ends; Europe's political actors generally accept the judgments of the ECtHR because they believe that they contribute to the protection of human rights, democracy, and the rule of law. A criticism of this Weberian approach is that despite people believing that the rules issued by authorities satisfy those ends, in reality, this may not be the case (Beetham 2013, p.11). Take for example the case of manipulation mentioned above. An authority can manipulate the people into believing that the way it exercises power contributes to the realisation of those ends while it does not. This is a textbook case of *abusing* power. Based on Beetham' criticism, a legitimate order is an order that subjects believe that it is well-ordered while *indeed* being well-ordered (*ibid.*, p.16).

I do agree with Beetham's criticism of the Weberian conception of legitimacy: AlgoAI engineers should always make sure that what they propose indeed contributes to the realisation of that order's ends. Otherwise, we have an abuse and/or misuse of power since they engineer a model that is expected to well-order an order but it fails to do. At the very best, it performs poorly, and at worst, it actually disorders the given order.

legitimacy & its *two* dimensions

freedom of expression. More on the interpretation and application of the law throughout the Thesis, especially in §II.4.1.2 where I show how *logic* can explicate them.

[27] E.g., British authorities and press have accused the Court for threatening UK's sovereignty due to its judgments on UK's treatment of prisoners' voting rights (*Greens and M.T. v. the UK (2010)*; *Hirst v. the UK (no 2) [GC] (2005)*) with the eventual execution of the judgments in 2017 being unreasonably delayed and inadequate (Nussberger 2020, pp.176-177).

Consequently, I adopt a two-dimensional legitimacy definition: (α) *ontic dimension*: an order *is* well-ordered; (β) *epistemic dimension*: subjects (and authorities) *believe* that the order is well-ordered. In other words, an order not only has to *be* legitimate, but it also needs to *look* legitimate.[28] We will see in §2.3 that according to Enlightenment's legitimacy paradigm, the belief to an order's well-orderness should be grounded on a ***rational justification*** *contra* other types of justifications like appealing to tradition or faith,

Nonetheless, normative accounts of legitimacy are still of relevance for descriptive accounts. Firstly, descriptive accounts of legitimacy depend *conceptually* on certain normative accounts. More precisely, a *de facto* legitimate authority is one whose rules are generally "*obeyed by subjects because many of them (or some important subset of them such as the officials of the state) think of it as having authority in the* ***normative*** *sense (Hart 1961).*" (Christiano 2020, §1, emphasis added).[29] For instance, in general, HCPs accept the authority of the ECtHR because *inter alia* they believe that the ends of the rule of law, human rights, and democracy are *normative*. Consequently, their decision to ratify the Convention is grounded on discourses about the normativity of those values. Secondly, many normative accounts of legitimacy like those of *consent* and *democratic approval* take into consideration what subjects actually believe (Peter 2017, §3.1,§3.3). There is one normative account though which does not do so and is of particular importance for the threat of algocarcy. It is the instrumentalist "*ends justify the means*" utilitarian account of *beneficial consequences* (*ibid.*, §3.2) that I will introduce in §2.8, ¶3.

Before concluding, I would like to justify my choice of definition for what a political authority is (§2.2, ¶¶2-3). As already stated, I do not intend to write a manifesto about how political orders should be oriented. I am rather interested about how political orders should *integrate* (or in the terminology of Danaher 2016 *accommodate*) AlgoAI to realise their chosen ends. Hence, I am interested in a *descriptive* account of political authority (what is called *de facto* political authority) and hence my choice of a *de facto* authority definition from Christiano 2020, §1. An alternative standard definition of *de facto* authority is one which identifies the concept of political authority with the concept of the Weberian legitimate political authority (*ibid.*). I rejected this option since I wanted to distinguish between subjects obeying an authority *via* false Legitimitätsglaube *contra via* true Legitimitätsglaube (§2.2, ¶9).

To conclude, for AlgoAI that exercises power in a political order to be legitimate, it needs to realise the ends of the said order's functional dimension (henceforth, I will call the operational requirements of those ends as *legitimacy requirements*) and that realisation should be *known* to the actors that make up that order. Things are not that simple though. Orders *overlap* and hence the same authority needs to realise ends from different orders which many times conflict with each other (*cf.* §1.1, ¶4). Even at the *same* order, political actors may disagree about how to realise those ends or even which those ends should be (remember the example of the *Perincek v. Switzerland (2015)* case in §1.1, ¶3). For instance, the Finnish government has to satisfy legitimacy requirements imposed by the national Finnish political order, the EU political order, NATO's political order, the CoE's political order, and so forth. Hence, AlgoAI engineers need to identify the legitimacy requirements imposed by each political order as well as their *relations* and *properties* (e.g., which requirements have *priority* over others in case of conflicts, which are obligatory, and so forth). I will call this problem the *overlapping orders problem*. It will be addressed throughout the Thesis (e.g., §2.4, ¶3; §2.6; [§III.]).

overlapping orders problem

After having introduced the concept of *legitimate order*, it is time to contextualise it in what Kissinger construes as the *Enlightenment era*.

I.2.3 On the age of weaponised reason

> "*The greatest problem for the human race, to the solution of which Nature drives man, is the achievement of a universal civic society which administers law among men.*"
>
> *Idea for a universal history from a cosmopolitan point of view*
> Immanuel Kant's 5th Thesis, 1784

Kissinger's ominous warning about Enlightenment's death is not just an eye-catcher. The value/fact dichotomy, the disciplinary partition of scientific practice, legitimacy, rights, liberalism, the separation of powers, SOCIETY 5.0, conceptual (re-)engineering, and many other concepts that have been or will be of relevance later on, all of them have their origins in Enlightenment's *dictum*: humanity *can* and *should* subjugate the natural and social orders *via reason* so as to *re*-order them towards specific *ends*.

[28] This position is on the same line of thought with jurisprudence's *dictum* "*justice must seen to be done*", a necessary requirement for an adequate realisation of the rule of law (Richardson Oakes and Davies 2016; *see also* https://www.venice.coe.int/WebForms/pages/?p=02_Rule_of_law (accessed May 10, 2023); "*jurisprudence*" is an alternative term for the *philosophy of law* (Leiter and Sevel 2022)). More on §2.4.

[29] Hart's "*The concept of law*" (1961) is a foundational, if not *the* foundational, book of contemporary jurisprudence.

More precisely, the *Age of Enlightenment* (*fr: Siècle des Lumières*; *de: Aufklärung*; henceforth simply *Enlightenment*) is an intellectual movement of the 17th and 18th century in Western Europe[30] whose central premise was the use of *human reason* to make sense of the world and ergo its alternative name as *Age of Reason*. *Interpreting* the world *via* reason is a *disruption* of the pre-Enlightenment interpretation of the world *via* religious *faith* (Bristow 2017, §1.2; Duignan 2023). I.e., certain propositions about the state of the world were justified *in virtue of* the *belief* of God making the world the way it is, while in the Enlightenment era, any propositions about the state of the world had to be grounded on a *rational* justification *contra non*-rational ones.[31] The pioneers of this intellectual movement are traditionally called *philosophes* (French for "*philosophers*") due to France being considered as the epicenter of the Enlightenment.[32] The views of the philosophes are diverse and many times conflicting: some advocated for enlightened despotism and others for democracy, some used reason to justify human equality and others like Kant to justify racial inferiority (Berlin 1993, p.27), some argued for humans being compassionate by nature while others for humans being guided by instincts of self-preservation (Munro 2021). Despite this diversity, one can identify positions that are more or less common throughout Enlightenment philosophes like the use of reason to interpret the world (Capaldi 1998, §1; Berlin 1993, §4).

We saw that in the pre-Enlightenment era God was used to justify why the world is the way it *is*. God was also used to justify *prescriptions* about how the world *should* be. And that included *justifications* about how *political orders* should be. Such a justification was the pre-Enlightenment legitimacy paradigm of the *divine right of kings* according to which "*kings derived their authority from God and could not, therefore, be held accountable for their actions by any earthly authority such as a parliament*" (Britannica 2021). Or John Locke's justification of the existence of *individual rights* in virtue of the relationship between God and humans (Capaldi 1998, p.351; *see also* Nickel 2021, §2.1). Enlightenment put an end to such God-driven interpretations of the world. The *new* political orders had to be grounded on *reason*.[33] In his "*Idea for a universal history from a cosmopolitan point of view*" Kant argued that the creation of a cosmopolitan order governed by the rule of law *via* reason is the "*greatest*", the "*most difficult*", and the "*last [problem] to be solved by mankind*" (*see* 5th and 6th theses in Kant 1963, pp.22-23).[34] For the philosophes, the formation of such an order is possible *inter alia* in virtue of two conditions: (α) a *mechanistic* conception of the social order; (β) the use of *reason* to rearrange the mechanics of social order so as to satisfy certain *ends*. More precisely, one of Enlightenment's revolutions was the mechanistic conception of nature that gave birth to many contemporary academic practices: *nature* is like a machine consisting of different *parts* and *relations* among those parts that *(causally) interact* with each other based on certain *(causal) laws* (Capaldi 1998, p.12). The quintessential paradigm of that revolution is Newton's "*Philosophiæ naturalis principia mathematica*" (*en:* "*The mathematical principles of natural philosophy*") (1687), the foundational work of contemporary physics and classical mechanics that includes the three laws of motion (Smith 2008, §2), i.e., causal laws that dictate *itner alia* how material bodies interact with each other *via* motion (Britannica 2023). Kant and other philosophes argued that the social world is of a *similar structure* (Capaldi 1998, p.8,pp.350-351). If one goes back to the definition of an order's *structural dimension* in §2.1, ¶2, the resemblance is pretty straightforward: an order constitutes of parts (social actors) *via* relations (social relations) that (*causally*) interact with each other *via* social actions.[35] In the next paragraph, we will see that the *functional*

origins of the *structural* dimension of social order

[30]Note though that MacIntyre, in his seminal "*After virtue*" where he criticises Enlightenment's failure to ground morality on *reason*, argues that Enlightenment is primarily *Northern European* (Scottish, English, German, and of course French) with other areas like Italy with the emblematic for Enlightenment Kingdom of Naples, Switzerland, South Germany, Austria, and Hungary, being mere "*outpost[s]*" of the Enlightenment culture (MacIntyre 2007, p.37). MacIntyre further argues that even French intellectuals who are traditionally considered to be the epicenter of Enlightenment's intelligentsia are in reality less influential for the movement than their Northerner Scottish and English counterparts (and of course Kant) (*ibid.*).

[31]I prefer using "*non-rational*" instead of "*irrational*" since the latter has a negative connotation and I do not want to prejudice the reader against the rejection of human reason as means to legitimise an order. As we will see, for some, that rejection is the right choice.

[32]Bristow 2017. Berlin 1993 and Capaldi 1998 use "*philosophes*" for Enlightenment intellectuals of every nationality, not *per se* French. I do the same because I intend to use "*philosophe*" as an umbrella term for all experts that partake and that should partake in the discourse about the new legitimacy paradigm (e.g., AI engineers, political scientists, logicians).

[33]The pre-Enlightenment legitimacy paradigm is not exhausted in the use of *faith* as a justification for how a political order should be. Another source can be e.g. political tradition (Kissinger, Schmidt, and Huttenlocher 2021, p.45; Hill 2010, p.123). For instance, in his seminal "*Towards perpetual piece*" (1795), Kant did not use "*the Treaty of Westphalia [a point of reference for the discipline of international relations (Kissinger 2014)] or any other supposedly foundational principle supplied by the past*" to make his arguments, but *solely* reason (Hill 2010, p.124). Regardless, *both* faith and tradition are about grounding legitimacy paradigms on something other than reason. In this Thesis though, the *non*-rational legitimacy paradigm of faith would be of particular importance and hence why I focus on faith instead of other non-rational legitimacy paradigms like tradition. Kissinger 2018 in his end-of-Enlightenment article does the same by comparing the "*Age of reason*" to the "*Age of Religion*" which it superseded.

[34]In the cited quotes, Kant does not use the concept of *political order*. It is Kissinger that makes this interpretation (2014, p.40). *See also* Hurrell 1990; Kleingeld 1998 for the same interpretation.

[35]Weber (1920) 1947 construes the interaction among actors as *causal* making his conceptual framework is another child of Enlightenment (Koshul 2005; Capaldi 1998, pp.19-20,306-308). This generalisation from natural sciences, what was called then *natural philosophy* (Janiak 2021, §1), is what gave birth to contemporary *higher level* disciplines that concern human activities: political science, social science, legal science, economics, analytic philosophy and its descendant, formal philosophy. The practices of those disciplines include the identification of parameters that influence other parameters (e.g., which parameters influence people's belief in the legitimacy of a political order

dimension of political orders is also rooted in Enlightenment.

A direct consequence of the mechanistic conception of social order is the position that by manipulating parts of that machine we can restructure it in different ways. If we identify how a part A (causally) influences part B then we can manipulate A to make B behave according to our desires so as to satisfy specific *ends*. We can become social *engineers* (Capaldi 1998, p.8,pp.19-20,pp.350-351,) like the then mechanical engineers that used the knowledge of the newly discovered laws of nature to design machines (e.g., Enlightenment's android automata of the harpsichord player "*La musicienne*")[36] leading eventually to the 1st industrial revolution (INDUSTRY 1.0) (Mokyr 2004, pp.34-35; *cf.* Voskuhl 2013, §6). At the same time, the philosophes used reason not only to identify how different parameters can be manipulated in order to organise the political order towards specific ends, but also to determine *which* are those ends. Once more, in the pre-Enlightenment paradigm, those ends were justified by appealing to God's wisdom like Locke's God-based justification of individual rights (*see* previous paragraph). In the Enlightenment legitimacy paradigm, the ends of a political order should be *rationalised*. This does not entail *per se* a rejection of God as was usually the case in the French philosophes (Bristow 2017, §2.3; Capaldi 1998, p.19-20). For instance, God can still be behind the existence of individual rights, but the philosophes want to justify them *rationally*. This rationalisation of ends is once more predicated on the Newtonian mechanical conception of nature, the so-called *naturalism* (Capaldi 1998, pp.11-12), like turning Locke's conception of rights "*into a quasi-Newtonian doctrine about the natural harmony of human interests*" (Capaldi 1998, p.351). Such a characteristic end is *happiness* (remember the US Declaration's "*pursuit of happiness*") with *utilitarianism* and *liberalism* being the quintessential ethical and political theories of Enlightenment respectively (pp.317,351; Bristow 2017, §2.1; *cf.* Cahoone 2023). Note that this does not entail that other political ideologies conflicting with liberalism do not belong to the Enlightenment paradigm. All big three ideologies of the 20th century fall under the paradigm of using human reason to engineer a social order towards different ends: Capaldi characterises *Marxists* as the "*most consistent and coherent representatives of the Enlightenment Project*" (1998, p.357) while Adorno and Horkheimer in their seminal "*Dialectic of Enlightenment*" published right after the atrocities of WWII (1947) interpreted Nazi death camps as "*what historically becomes of the supremacy of instrumental reason asserted in the Enlightenment*" (Bristow 2017, §2.1). One can propose to reduce crime rates by resolving to eugenics or by adding more psychologists to educational facilities or by introducing harsher punishments and more police or by creating more job opportunities and enhancing social mobility (*cf.* Capaldi 1998, p.8). All those are attempts to *rationally* justify how ends found in the Newtonian nature can be realised.

origins of the *functional* dimension of social order

Even if one appeals to the mechanistic conception of nature to justify the choice of moral & political ends (e.g., nature bestowed us with certain rights, the so-called *natural rights* which were precursors of the post-WWII concept of human rights),[37] one still has to *rationally* justify how we *discovered* those natural ends. Resolving to God in order to justify them is no longer possible. For instance, two strands of the rational justification of the Enlightenment Project are *rationalism* and *empiricism* (Bristow 2017, §2.2). According to rationalism, moral and political ends should be *deductively* derived by axioms that we already know intuitively (Markie and Folescu 2023, §1.1). A prime example is the Amsterdam-born Dutch philosopher (and philosophe) Baruch Spinoza (1632-1677) who in his posthumous "*Ethics: Demonstrated in geometrical order*" (*latin:* "*Ethica ordine geometrico demonstrata*") tried to ground his moral and political theory on an axiomatic deductive system similar to that of Euclid's Elements using axioms like "*Everything which exists, exists either in itself or in something else.*" and "*That which cannot be conceived through anything else must be conceived through itself.*" (Axioms I & II respectively from Part I in Spinoza 2017; *see also* Nadler 2022, §§1-2; Steinberg 2022). In juxtaposition to rationalism, the justification of ends in the strand of empiricism does not start *deductively* form the *mind* but *inductively* from our *empirical* experiences with nature itself (Bristow 2017, §1.2). A founding figure of the empiricist strand is Third Earl of Shaftesbury with his "*Characteristics of men, manners, opinions, times*" (1711), in which he argued *inter alia* that we can understand what is moral or not by reflecting on our actions. For instance, when we reflect on actions of gratitude or kindness we find ourselves liking them indicating their morality, while when we reflect on actions of jealousy or resentment we find ourselves disliking them indicating their immorality. Such *empirically* grounded reasoning can guide us to discern different moral and immoral actions.

rationally justifying ends

This rational Newtonian construal of moral and political values presupposes an *subject independent order of values* (henceforth *ordo essendi* (Iatrou 2022, §2; *cf.* de Jong and Betti 2010; Cantù 2014) that is *epistemically accessible* by human reason. In other words, it presupposes *objectivity*: values are objectives in the same manner that physical objects are and are governed by (causal) laws in the similar way that the natural world is (Capaldi 1998, p.12; Bristow 2017). As we saw in §1, the objectivity of values is currently challenged both on pragmatic

(Beetham 2013, pp.8-9; Weber 1974, pp.78-79; 1947, pp.130-131; Peter 2017, §1)) due to the *(causal) laws* of the social order. (Capaldi 1998) (*cf.* Brown, McLean, and McMillan 2018, p.966).

[36] Voskuhl 2013, pp.2–3. Enlightenment androirds are premised on the position that human as part of nature is a *machine* (*ibid.*). A seminal piece of Enlightenment thought is de la Mettrie's "*L'homme machine*" (*en:* "*Man a machine*") (1747).

[37] John 2011, pp.33,198; Gray 1995, p.235. *See e.g.* Spinoza's and Hobbe's conception of *natural rights* (Steinberg 2022, §2.1).

and philosophical grounds. This challenge once more originates from Enlightenment. For instance, the Scottish philsophe David Hume (1711-1776)[38] rejected the possibility to epistemically accessing an *ordo essendi* of values and he argued that morality is grounded on our subjective feelings or attitudes towards what we consider to be moral. Hume is considered to be a founding figure of *ethical subjectivism* (Bristow 2017, §2.2) what came to be known today as *non-cognitivist* ethics (van Roojen 2018; Shecaira 2011).[39] Hume's position exchanged his burden of "*explaining how the objective order of values belongs to the natural world as it is being reconceived by natural science*" with the burden of "*explaining how error and disagreement in moral judgments and evaluations are possible*" (Bristow 2017, §2.2). I.e., Hume had to faced the objectivity challenge.

ordo essendi

One of Hume's arguments in favour of subjectivism was the *ought/is* dichotomy (aka *Hume's Law*) which is the precursor of the fact/value dichotomy. Hume argued that one can not derive an *ought* from an *is*. E.g., from the *is*-statement "*For you to do X under circumstances Y* ***is*** *good.*" one can not derive the *ought*-statement "*You* ***ought*** *to do X under circumstances Y.*" unless one includes in their premisses an extra *ought*-judgment like "*For any possible circumstances, you* ***ought*** *to do whatever is good under those circumstances.*". In other words, you need an *ought* to derive an *ought* (Putnam 2002, pp.14-16). From a logical point of view, this motivated the introduction of non-traditional logics to deal with inferences of ought statements with the classical example being von Wright's deontic logic[40] (Habermas 1992, p.102; in this book, Habermas examines the implications of Hume's ought/is a dichotomy to the problem of objective truth in justifications of legitimacy). According to Putnam's interpretation of the dichotomy,[41] Hume argument is not based on the different *logical forms* of *ought*- and *is*-statements. It is rather a *metaphysical* argument. Specifically, *is*-judgements are judgements about what Hume calls *matters of fact* while *ought*-judgements are judgements about *ideas* which are metaphysically distinct from the factual natural world. Hence, one can not use factual knowledge to derive truths about ideas.[42]

origins of the fact/value dichotomy

To sum up, for Enlightenment's philosophes, reason can be used to identify certain ends. Those rationally justified ends show what *ought* to be the case. Humans ought to be free, self-governed, to live and to live with dignity, and so on. And the philosophes *acted* upon those imperatives. *Contra* to contemporary academic philosophical practice, Enlightenment was an era during which "*philosophy did constitute a central form of* ***social activity***" (MacIntyre 2007, p.36; emphasis added). The ideas of the philosophes were spread through the use of the technology of the *printing press* (Sunder 2020, p.999; Burrows 2015). Reason became "*armed*" resulting in the overthrowing of the old political orders (e.g., the so-called *ancien régime* in France) establishing new ones *ad initio* (Kissinger, Schmidt, and Huttenlocher 2021, p.45). French Revolution was the result of the writings of philosophes like Rousseau, Montesquieu, Voltaire, and Diderot. The new French order was based on values popularised by the philosophes like *consent* of the governed and *self-governance*, the *separation of powers*, natural rights like *freedom* and *equality* (Kumar 2020; Burrows 2015; D'Agostino, Gaus, and Thrasher 2021, §3.1), values that motivated the revolutionaries to draft the *Déclaration des droits de l'homme et du citoyen du 1789* (*DDHC*, *en: Declaration of the Rights of Man and the Citizen*) and the 1791 French constitution, to establish *distinct* legislative and executive authorities that were staffed by *elected* representatives (Crook 2015; Fitzsimmons 2015; Edelstein 2014). Similarly, the American Revolution and the founding of the USA were a result of the writings and social activism of Thomas Paine, James Madison, Thomas Jefferson, John Adams, Benjamin Franklin, and other philosophes.[43] In other words, what *ought* to be the case in the writings of the philosophes, *became* the case in the actual social order; the *ought* turned into an *is* and reason was the *weapon* that executed this reordering by *rationally* identifying the *means* to do so.

turning the *ought* to *is*: the *weaponisation* of reason

Law is the quintessential *rational means* to perform such a reordering. If we can rationally identify ways of restructuring a social order so that everyone is free, happy, and equal, then we ought to enshrine those proposals in our laws like the French and the Americans did with their constitutions or like Kant envisaged with the "*perfectly just civic constitution*" of his ideal cosmopolitan order (Kant 1963, 5th Thesis). As Kissinger argues, Enlightenment gave birth to the ideal of a "*reasoned*" "*rule-bound*" international order (Kissinger, Schmidt, and

rule of law's ontological priority

[38] For the parallels between Hume's philosophy and Newton's natural philosophy *see* Schliesser and Demeter 2020 or Capaldi's 1975 "*David Hume: The Newtonian philopher*".

[39] Capaldi 1998 and MacIntyre 2007 (p.14) place Hume in the origins of the *emotivist* strand of non-cognitivism.

[40] von Wright 1951; *cf.* §II.4.1.2. *See* Hilpinen and McNamara 2021 for a philosophical and historical introduction to contemporary formal deontic logic.

[41] Despite the ought/is dichotomy being the content of only a single paragraph of Hume's seminal "*A treatise of human nature*" (1739-1740, Book III, Part I, Section I, ¶27), it is the subject of intense interpretational controversy (Cohon 2018, §5). Putnam explicitly distances himself from many of those interpretations providing another interpretation in this mosaic (2002, p.14).

[42] What I construe as *factual judgements* is different from Hume's matters of fact which are a subset of the former. More precisely, in §1.1.1, ¶8, I construed judgements about relations of abstract entities like "$1 + 1 = 2$" as factual judgements while Hume considers them to be judgments about ideas and hence distinct from matters of fact (Morris and Brown 2022, §7.1). The fact that they are judgements of ideas does not entail that they are ought-judgements like evaluative judgements are. They still describe what *is* the case, although they do so for abstract entities, and hence my choice of classifying them as factual judgements.

[43] Colbourn 1998, §Part II; Ralston, n.d.; Foner 2005; Sunder 2020, pp.997-1001; §2.5, ¶12.

Huttenlocher 2021).[44] The rules of this rule-bound order are the *laws* that allow for the ends of the functional dimension to be realised. What makes laws quintessential is *inter alia* their *ontological priority* over the realisation of an order's ends: we *first* introduce the laws that order an order and then we order the order based on those laws so as to realise the desired ends. Remember Huntington's 1968 quote at the very beginning of this section (§2): the value of liberty is realised *after* an order is ordered appropriately. The ontological priority of laws is also explained by the *mechanistic conception* of social orders: engineers first *design* blueprints (drafting laws), then they *build* a machine based on those blueprints (re-ordering the social order), and then the machine produces the desired outcome (the realisation of the social order's functional dimension).[45] This ontological priority is what makes the value of the *rule of law* universal across different Enlightenment legitimacy paradigms at least in its minimal form: *everyone*, even those that exercise power, *should* abide by the law. Hence the *maxim* "*government by law and not by men*" (Raz 1979, p.212). For an order to be well-ordered, realising this dictum is a necessity. Hence our first legitimacy requirement: adequately realising the value of the *the rule of law*.

I.2.4 On the rule of law

Joseph Raz's minimal definition of the rule of law in his seminal "*The authority of law*" is a minimal definition that everyone would accept: "*'The rule of law' means literally what it says: the rule of the law. Taken in its broadest sense this means that people should obey the law and be ruled by it. But in political and legal theory it has come to be read in a narrower sense, that* ***the government shall be ruled by the law and subject to it.*** *The ideal of the rule of law in this sense is often expressed by the phrase '****government by law and not by men****'.*" (p.212, emphasis added). Both government and citizens, all ends of political relations, are expected to be *ruled* by law; law *is* the highest authority. This *supremacy of law* is referred to as the value of *legality*:

the value of *legality*

> "'*[Legality (supremacy of the law)] first implies that the law must be followed. This requirement applies not only to individuals, but also to authorities, public and private.*"
>
> REPORT: *On the rule of law (CDL-AD(2011)003rev), ¶42*
> Venice Commission, Strasbourg, 4 April 2011

The controversy regarding rule of law's definition starts when one has to argue whether the rule of law entails obligation about the *content* of the law: does the rule of law entail that human rights or fair election shall be protected by the law? A content-free conception of law is traditionally referred to as a *formal* conception of the rule of law *contra substantive* conceptions (Nishigai 2021, p.495; this distinction is attributed to Raz 1979, §11). It is quite common for authoritative authorities to adopt a formal conception of the rule of law, arguing that for their authority to be legitimate according to the rule fo law, all that they need to do is follow the law they themselves establish (CDL-AD(2011)003rev, ¶15; CDL-AD(2016)007, ¶12).

formal v. substantive legitimacy requirements

For the CoE, realising adequately the rule of law *does* entail that the law should have specific *substance* (CDL-AD(2011)003rev, ¶15). And that substance *does* include realising adequately the values of human rights and democracy. In reality, all three pillars constitute a coherent whole where each pillar depends conceptually on the other two; none of the three pillars can be realised adequately without the others (CDL-AD(2016)007, ¶12). For instance, regarding the intertwining of rule of law and human rights, Venice's Commission's operational definition of the rule of law includes the operational requirements that national constitutions and legislations should include the prohibition of discrimination based on gender, language, political opinions, and other status, as well as that they should ensure that the law will treat similar situations similarly and dissimilar situations dissimilarly (the so-called *principle of equality*) (CDL-AD(2016)007, p.18). The rights to prohibition of discrimination and equal treatment before the law are *human* rights protected by the Convention (*see e.g.* ARTICLES 6 & 14). Ergo, the Commission's definition of the rule of law requires that the law of a state shall protect certain human rights and that by protecting those human rights the state protects the rule of law. Similarly, regarding the intertwining of rule of law and democracy, Venice's Commission's operational definition of legality includes the operational requirements of the elected parliament being supreme in deciding the content of the law, the parliament providing adequately justified explanations about any proposed legislation, as well as the parliament debating such legislation publicly (*ibid.*, p.13). In this case, the Commission's definition of the rule of law requires that the law of a state should protect certain democratic principles and procedures and that by doing so the state protects the rule of law.

European order's *substance*

Let's go one step back and look again at legitimacy from a *formal* perspective before deciding to put any substance to it. We have seen that legitimacy's ontic dimension requires an order to be well-ordered and that an

[44] *See e.g.* ¶2 of the Declaration (emphasis added): "*...whenever any Form of Government becomes destructive of these ends [life, liberty, pursuit of happiness, etc], it is the Right of the People to alter or to abolish it, and to institute new Government,* ***laying its Foundation on such Principles, and organizing its Powers in such Form****, as to them shall seem most likely to effect their Safety and Happiness.*".

[45] *Designing* and *building* are two of the three phases of the practice of *engineering*. More on §II.4.1.1.

order is well-ordered *only if* it realises the value of legality. Subsequently, the value of legality entails the legitimacy requirement that all *legal* requirements for an AlgoAI model are also *legitimacy* requirements: AlgoAI models need to abide by the existing law. This conclusion provides a first solution to the problem of overlapping orders. More precisely, legitimacy requirements can be construed as *norms* since they describe what *must* or what *can* be the case (modalities of *obligation* and *permissibility* respectively; in logic, both modalities are traditionally construed as *norms* (Hilpinen and McNamara 2021, §§4-5)). Ergo, in case of conflict among legitimacy norms, the principle of legality forces us to prioritise *legal* norms over other norms like ethical or customary norms. At the same time, in case of conflicts among legal norms, the principle of legality forces us to prioritise legal norms based on principles of prioritisation prescribed by the law itself. Specifically, a quintessential characteristic of legal reasoning is *defeasibility*. I.e., there are often cases where multiple conflicting norms are applicable, and hence, the legal expert needs to determine which norm prevails and which are *defeated* (Governatori, Rotolo, and Sartor 2021, pp.688; *cf.* Poggi 2021; Hage 2005, §1). The law usually provides criteria to decide which norms are defeated. Three traditional criteria for making such a decision are the *lex specialis* principle where the more specific norm prevails (e.g., in the *specific* case of armed conflict the *general* human rights laws like those of the Convention are defeated by the more specific *law of armed conflict* (Chevalier-Watts 2010; *cf.* §2.7)), the *lex superior* principle where the norm issued by the actor with the highest authority prevails (e.g., the norms of the Convention or of the ECtHR's case-law prevail over the HCPs' constitutional courts which in their turn prevail over the national non-constitutional law), and the *lex posterior* principle where the most recent norm prevails (e.g., the ECtHR should use in its judgements norms derived from the most recent interpretations of the Convention which many times contradict norms derived from the older interpretations (Dzehtsiarou 2011, 1731; *cf.* §§II.4.1.2,III.3.2.2))[46] (Governatori, Rotolo, and Sartor 2021, p.689). I will call the hierarchy of norms induced by the principle of legality as the *legal order* of the legitimacy requirements. This priority of the legal order of norms also entails that AlgoAI engineers should use *legal* concepts in their practice and not their *non*-legal counterparts. I.e., we are concerned with the *legal ordo essendi* and not the political, ethical, etc ones (*cf.* §2.7, ¶1; §1.1, ¶2).

legal requirements *are legitimacy* requirements

legal order

In the previous paragraph, I argued that AlgoAI models need to "*abide by the existing law*". Much of the already existing law is what is called law 1.0, i.e., the traditional OG law where a set of norms regulates human social activity (Brownsword 2021, §I.3). AlgoAI though is a *disruption* of human activity; there was no ChatGPT to write judgements about Colombian insurance law one year ago (L. Taylor 2023). Therefore, we should not expect law 1.0 to be capable of providing an adequate list of legitimacy requirements for concepts that were not existing at the time it was introduced. The drafters of the Convention were not aware of concepts like large language models, phishing, defeasible logic programming, or IP address when they were drafting the Convention. Hence, law 1.0 should both be *adjusted* as well as *expanded* to regulate challenges emerging from new technologies (e.g., designate drone exclusion areas) (Brownsword 2021, §I.3). That new law is called law 2.0 (*ibid.*, §I.5). We have already seen examples of law 2.0 (the laws against revenge porn or about the taxation of cryptocurrency transactions in §2.1, ¶7), and we will see any more examples later on as well as their expansion to laws 3.0 & 4.0. Any *new* law like law 2.0 should be *oriented* towards the ends of the political order's functional dimension. Otherwise, we are disordering that order and ergo we are delegitimising it. Furthermore, legality also entails that any new law like law 2.0 should be introduced according to the *procedures* described by the already existing law. It is the law that dictates how law-making should be done (CDL-AD(2016)007, pp.11,13).

laws 1.0 & 2.0

The requirements laid out so far were about the *ontic* dimension of legitimacy. Legality entails that that law must be applied adequately, i.e., it dictates what should *be* the case. We also need to make sure that the subjects have *adequate epistemic access* to the adequate application of the law. For Weber, subjects *knowing* that power is performed according to procedures prescribed by the law is the most common contemporary source of legitimacy (Weber 1947, p.131). Consequently, courts should provide *public justifications* for their judgements, those judgements should be written in *plain* language, they should be *affordable* if not free, *translated* to other langauge and *braille* when requested (never forget *sign languages* as well), and everyone should have access to them (e.g., if judgements are provided online, there should be appropriate support for citizens with no internet access or with digital literacy difficulties). WJP classifies such requirements under the concept of *open government* (WJP 2022, pp.14-17; *cf.* article 6 (right to a fair trial), ¶3).

the value of *open government*

Open government does not suffice for an adequate realisation of legitimacy's epistemic dimension. The belief that the law is applied adequately is not only about how authorities have acted, but it is also about the *expectation* that authorities *will* act similarly in the future. This requirement is subsumed under the principle of *legal certainty* (Fenwick and Wrbka 2016). Open government corroborates legal certainty: knowing how the government operates grounds expectations about its future behavior (CDL-AD(2016)007, pp.15-17). Another essential requirement for legal certainty is *forseeablity*: it should be foreseeable how the law is to be applied (CDL-

the values of *legal certainty & foreseeability*

[46] For a specific example, *see* the conflict between the controversial *Osman v. UK (1998)* case and the *Z. & others v. UK (2001)* case in Nolan 2013, p.288.

AD(2016)007, p.15). In many cases, the ECtHR deems that an actor that has violated human rights should not be held legally responsible if it was not foreseeable that they were violating the law (e.g., ECtHR Registry 2021, §III.B.1). Finally, legal certainty requires specific *logical* properties of the argumentative structure of the justifications judicial authorities provide to justify their judgements. E.g., are the arguments coherent (Letwin 2021)? Are there contradictions (*ibid.*)? If there are exceptions to the law, are they sufficiently justified (e.g., by appealing to the *lex superior* principle) (CDL-AD(2016)007, p.16)? Is there ambiguity or vagueness (Raz 1979, p.214)? Interestingly, the Enlightenment-based legitimacy requirement of the *rational* justification of judgments and its *public* character is reflected in the change of *architecture* of French courtrooms during and after the Revolution (K. F. Taylor 2013).

Note that legal certainty is fundamental for *both* epistemic and ontic dimensions of legitimacy. *Contra* to the physical norms of the natural order, the legal norms of the social order do not *oblige* the objects whose behavior they regulate to act according to their content. The implementation of the laws of a political order depends on how the political actors *interpret* those laws and how they decide to **act** upon their interpretation . If there is no certainty about the content of the law, then the subjects will not act as expected and the order can not become well-ordered. Ergo, the lack of legal certainty undermines the *ontic* dimension of legitimacy.

The foregoing rule of law requirements are also of particular importance for the legitimacy of *judicial authorities*. Apart from imposing obligations to a court's judgements (e.g., the judgments being *just* & adequately *justified* without *contradictions* and *uncertainty*), they also impose *procedural* obligations (e.g., there should be procedures that allow subjects to request information about a judgement or to object to a given judgement; *cf.* ARTICLE 6 (RIGHT TO A FAIR TRIAL), ¶3; WJP 2022, p.17), as well as obligations about the *qualities* of authority-actors. For instance, judges should be **competent** enough to deliver just judgements, accompanied with adequate justifications written in plain language and in a timely manner (i.e., they should not *misuse* power) (*cf.* Spaak 2009; WJP 2022, pp.14,16). They should also act *ethically* like acting impartially and denying bribery (i.e., they should not *abuse* power) (*ibid.*, pp.15-19]). Finally, it should be noted that all proposed rule of law requirements, even open government, have *not* been justified on any request for substance like human rights or democracy. Therefore, they are not substantive, but *formal* requirements for *any* political order. This conclusion is on par with Raz's rule of law formal concept in Raz 1979, pp.214-219 (*cf.* §2.4, ¶1) as well as with rule of law's ontological priority (§2.3, ¶9)]. Having said that, as I will argue in the rest of §2, there can be variations in certain qualities of those formal requirements depending on the substance of each political order. E.g., we will see in §2.8, ¶2 that for illiberal democracies, well-orderness does not contradict *per se* with restrictions to legal certainty. At the same time, there still needs to be at least a sufficient level of legal certainty so as to order democratic political orders illiberally. Considering this, it may be more appropriate to distinguish between *weak* and *strong* formal requirements with legality being a strong formal requirement and legal certainty being a weak formal requirement.

rule of law & judicial authorities

weak v. strong formal requirements

Before concluding, I would like to provide additional arguments of why *epistemic accessibility* contributes to the legitimacy of judicial authorities with regards to the *separation of powers*. I do so because the same arguments will be used to deal with the threat of algocracy. To make my case, I will contrast judicial authorities with *elected* by the public authorities (e.g., parliaments, heads of states, etc.). In the case of elected authorities, the subject directly votes for those who believe that will exercise power legitimately. They also know that they can vote them out in case that they are not satisfied with the way they exercise power. I.e., elections make those authorities *answerable* to the public. Judicial authorities though *are* and *should* remain *unanswerable* to circumstantial majorities in order to be able to apply the law to those majorities.[47] They should also remain *uninfluenced* by the general public's opinion since what is *just* is not *per se* popular.[48] In the same line of thought, judicial authorities judge whether authorities elected by majorities act according to the law. For instance, judicial authorities can restrict the acts of elected authorities if they violate the constitution like trumping legislation voted by parliamentary majorities.[49] In other words, many times, judicial authorities have to act against what the majority *believes* should be the case; they have to be *anti-majoritarian* (or *counter-majoritarian*) (Robertson

epistemic accessibility & check-and-balancing judicial authorities

[47] This is a strong formal requirement. It is an implementation of the "*government by law and not by men*" *dictum*. Note that unanswerability to the general public does not *per se* decrease the public's Legitimitätsglaube. As Vibert remarks, while the majority of the public may disagree with particular judgements, what is needed is a *general* acceptance of a judicial authority's judgements. And indeed, experimental evidence shows that the public can have such general acceptance despite strong diagreements with particular cases (Vibert 2007, p.115-116; [more]).

[48] This requirement is premised on the argument that an opinion about the interpretation and application of the law that is supported by a majority is not *per se* the *just* way to interpret and apply the law. Classical examples are the trial of Socrates, Jesus, Emperor Marcus Aurelius (Mill 1901, pp.44-50), and the anti-Semitic Dreyfus affair case. Consequently, for the law to be applied adequately and not be dictated by the so-called *mob rule*, the court should judge independently (Vibert 2007, pp.115-116). However, we will see in §2.8, ¶2 that there are political orders like illiberal democracies where the opinion of the majority should have substantial weight in judicial decisions. Ergo, independence from the public opinion is a *weak* formal requirement.

[49] This is an exercise of the so-called *judicial review* power, the power of judicial authorities to judge whether authorities act according to the law (Tate 2023; Fordham 2020, p.5). I.e., judicial review is essential for *legality*: "*Judicial review is the role which the Courts have*

2004, pp.49-50) And at the same time, that majority does not have the ability to make them answerable.[50] It is also judicial authorities that judge whether the judicial authorities apply the law(!) It is therefore of prime importance for the separation of powers that judicial authorities *justify* their political activity with *public* and *easily accessible* (e.g., affordable, written in plain language, translated, etc) justifications. Due to a lack of alternatives, forcing judicial authorities to justify to that extent their actions is a safeguard that they will not abuse/misuse power. Furthermore, those justifications can be used by judicial authorities to check-and-balance other judicial authorities (e.g., if a judgement is problematic then a higher court may overturn it). They can also be used by the public and non-judicial authorities to legitimise an interference with judicial independence under exceptional circumstances. E.g., citisens proposing legislation to replace human judges with robot judges in order to deal with gruesome gender bias in judgement about sexual abuse cases (*see e.g.* §3.2.1, ¶1; this is a check-and-balance against *misuse* of power) or executive authorities investigating judges for bribery, conflict of interests or other types of power *abuse*. Note that as argued in §2.2, ¶9, in Enlightenment's legitimacy paradigm, those justifications should be grounded on *human reason*. I.e., they should have the logical form of human reasoning methods like deduction, case-based reasoning (*fn.* 24), causal inferences, or other judicial reasoning methods (more details on judicial reaosning methods along with a list of citations on §II.4.1.2). The position that judicial authorities should not be answerable to the public while still providing justifications for their judgements based on specific justificatory principles was also a position that the Enlightenment American philosophes had when drafting the US' first constitution (Vibert 2007, p.171).

In an algocratic order, judicial authorities are not the only *unelected* authorities. We vote neither about which ALGOAI models should contribute to the exercise of power nor we vote about which ALGOAI engineers should engineer those models, while as we will see in the next subsection *both* ALGOAI models and ALGOAI engineers *co*-exercise power. Considering this, in §2.5, I elaborate on the details of that co-exercise of power as well as which should be the role of ALGOAI engineers in the separation of powers of a *legitimate* algocracy according to the *rule of law*. From this *new separation of powers*, we will derive new rule of law legitimacy requirements for both ALGOAI & its engineers that will pave a way out of the threat of algocracy.

I.2.5 On the fourth power: unelected epistemic authorities

In today's world, there are more types of political power than the traditional three branches depending on the particularities of each political order. For instance, former judge Stephen Sedley (served at both British courts and the ECtHR) argues that in contemporary Britain there are at least three new branches: the church, the media, and the security & intelligence services (Sedley 2015, p.190-192, *cf.* Beatson 2021, §8.I). Note that these examples include *non-state* actors that exercise power (e.g., privately owned corporate media or YOUTUBE channels). At the same time, the *strength* of a power branch differs depending on the particularities of each political order. For instance, the church has less influence in France's political order due to the so-called value of *laïcité* (*see* Article 1 of the French constitution; Sedley 2015, p.191), *contra* Iran's theocratic state (EIU 2023, p.63) or *contra* other political orders like Greece's which is situated in-between those the two extremes of the spectrum (Kaltsas et al. 2022).

non-traditional types of power

From the non-traditional types of political power, one is of particular importance for the threat of algocracy, those that Vibert calls "*the unelected*" (henceforth the *fourth* political power). More precisely, in his 2007 "*The rise of the unelected*", Vibert argues that the bodies of *unelected experts* established by governments with the purpose of *gathering* and *processing specialised* information so as to exercise power should be construed as a separate branch of power (*ibid.*, pp.2,30-33). The reason for acknowledging them as political authorities is to impose appropriate checks and balances (§2.2, ¶5). Such unelected authorities are (inter)national committees on food and environmental safety, independent central banks and broadcasters services like the BBC, bureaus of statistics, bodies of space, marine life or meteorology research, organisations about humanitarian aid, trade, or labour rights, and so on. The IMF, the WHO, the OECD, the UN, the EU, the CoE and most of their organs are such international bodies of unelected (*ibid.*, pp.19-30, 144-148, §9). Examples of such unelected bodies used in this Thesis are CoE's Venice Commission, CEPEJ, MSI-AUT, & Committee on Artificial Intelligence (CAI), UN's Civil Society Unit, EU's PEGA, JURI (*see* its study on ALGOAI that regulates free speech: Sartor and Loreggia 2020), eu-LISA & Eurojust (*see* their 2022 joined report "*Artificial intelligence supporting cross-border*

the *fourth* power

established for upholding and enforcing the ***rule of law*** *in the context of public authorities. It ensures that public authorities are accountable to law, securing that their public functions are undertaken according to law. It means, in a practical and effective way, that public authorities are not "above the law'."* (*ibid.*, p.8; emphasis added). Once more, in certain political orders like illiberal democracies judicial review is restricted when it comes to check-and-balancing elected authorities (*cf. fn.* 48; §2.8, ¶2).

[50] In many cases though, elected authorities participate in the procedures of deciding who will become a judicial authority-actor like in the case of selecting judges for the ECtHR: "*The judges shall be elected by the* ***Parliamentary Assembly [PACE]*** *with respect to each High Contracting Party by a* ***majority of votes*** *cast from a list of three candidates nominated by the* ***High Contracting Party***" (ARTICLE 22 (ELECTION OF JUDGES), emphasis added; *see also* Lemmens 2015; Kosa 2015).

cooperation in criminal justice").

Note that based on the definition of power laid out in §2.2, it is not necessary for the unelected bodies to be the final decision-makers in order to exercise power. More precisely, *advisory* bodies that provide *factual support* in favour of particular political actions performed by the traditional authorities are also compatible with the definition of power. Take the example of the WHO or of the national unelected bodies of medical experts that shaped the (inter)national health policies during the pandemic (Singh et al. 2021 and Akhtar 2022 respectively). It may have been the case that it was the governments and not those unelected bodies that made the final decisions regarding which policies should be adopted like what happened in Norway (Christensen and Lægreid 2020, p.778). However, governments adopted policies, policies with rather undesirable consequences like harming the economy (*see e.g. ibid.*, pp.775-776; Akhtar 2022, pp.256-258), in virtue of the factual support provided by the experts (Bylund and Packard 2021, §3). Similarly, this factual support contributed decisively to the public believing that they ought to follow the proposed measures regardless of their undesirability: "*...trust in the credibility of public health experts, health systems and scientific evidence has been shown to encourage[...] compliance with lockdowns, and adoption of preventive public health measures, such as physical distancing and mask-wearing, all of which are difficult and costly to implement without public support and commitment.*" (Lazarus et al. 2020, p.12). If exercising power is among others *making subjects generally obey rules even if they do not want to obey them* (§2.2, ¶¶4,6), those examples are textbook cases of exercising power. Authorities that exercise this type of power are known as *expert* or *epistemic* authorities (Liese et al. 2021, p.356).

epistemic/expert authority

This distinction between providing *factual evidence* and performing *evaluative judgements* based on those factual judgements is the essence of the proposed separation of power:

> "*What underlies the new separation of powers is a distinction between the empirical component of public policy and the value judgements. The making of public policy involves both elements – the factual evidence and the social or political judgements to be made in the light of that evidence. Unelected bodies have an advantage in dealing with the empirical components of public policy and elected bodies in choosing the values to be reflected in public policy.*" Vibert 2007, p.2[51]

factual but *not* evaluative judgements

In other words, the normative is for experts to *advise* the traditional authorities on how to reorder the social order towards specific ends based on rationally justified factual judgments. But it is still those traditional authorities that will decide *which* those ends should be as well as *how* the factual judgements of the experts will be used to realise those ends. This is nothing more than an *institutionalisation* of the philsophes like law being an institutionalisation of ethical and political philosophical positions (§1.1, ¶2). Kant, in his seminal "*Towards perpetual peace*", made a similar proposal for the states to consult the "*maxims of the philosophers*" in case of war (Kant, p.93 2006), a proposal that Kissinger considers a paradigmatic example of Enlightenment's weaponisation of reason in order to re-order a political order (Kissinger, Schmidt, and Huttenlocher 2021, p.45). In his proposal, Kant clearly advocates for the role of the philosophes to be *advisory*: "*...I do not mean to say that the state must favor the principles of the philosopher over the pronouncements of the lawyer (as a representative of state authority), but rather only that one* ***listen*** *to the philosopher.*" (Kant 2006, p.93). Since it is the traditional authorities that are legitimised to perform the evaluative judgements, allowing them to differ from the advice of the experts can boost the trust of the public in the policies adopted after consultation from the said experts. It is still the legitimate authorities that are already trusted by the public that are in control of the political order's structure. Indeed, during the pandemic, the strong collaboration of the Norwegian government with the national medical authorities in which the government retained its power to make the important decisions and ergo diverging from the experts' advice is credited as one of the reasons for the corroboration of the Norwegian political order's legitimacy (e.g., citizens' satisfaction with the democracy increased from 57% to 72%) as well as on of the reasons for the success of Norway's response to the pandemic (Christensen and Lægreid 2020).

Vibert in the foregoing quote argues that it is the *elected* authorities that should perform value judgements, while I generalised their position to "*traditional authorities*" (executive, legislative, judicial), elected or not. I did so because whatever the type of regime or political power, if experts are assigned the task of policy-making in the place of traditional authorities, then, by default, they replace those authorities in the exercise of power. Ergo, we should make sure that the replacement is done in a *legitimate* way no matter if they are elected or not. Otherwise, we have an illegitimate exercise of power both from the experts and the authority that transferred them their power. More precisely, according to the value of legality, for a transfer of power to be legitimate, it should be performed as prescribed by law. If the law does not specify how power should be exercised then the law can not govern the authority that exercises that power leaving room for illegitimate exercise of power. Consequently, if there is no such law, then such law should be introduced (LAW 2.0) and its introduction should once again be in accordance with the already existing laws (LAW 1.0). What also matters in terms of legitimacy is which is the *legal source* that establishes the separation powers. Is it the constitution, the case-law, or non-

the new separation of powers & *legitimacy*

[51] "*Public policy*" refers to the acts performed by state-actors to deal with specific problems (e.g., a pandemic) (Knill and Tosun 2012, p.4).

binding directives? Violating the constitution undermines legitimacy more severly than what violating non-binding directives does (*cf.* (CDL-AD(2016)007, p.17)). An example of an adequate *constitutional* separation of the four powers is Sweden, a separation that restricted the illegitimate use of power during the pandemic by all four types of authority more adequately than other political orders. Note that in Sweden, expert authorities are not merely advisory bodies, but they also have the independence to enact policies (executive power), albeit there are restrictions to keep such policy enactments within the scope of factual evidence-based policy-making (Bylund and Packard 2021).

check-and-balancing epistemic authorities

How the fourth power should be checked-and-balanced in this new separation of powers? Similarly to judicial authorities, epistemic authorities should be *uninfluenced* by and *unanswerable* to the opinions of majorities. The truth of factual judgements (e.g., the truth of the proposition "*Wearing masks in public reduces the transmission of the virus.*") is not to be decided by vote, opinion polls or any public pressure. Neither should the experts be elected by the public opening the doors for scientific populism. At the same time, epistemic authorities have the power to delegitimise the policies of elected authorities by arguing that they are factually ill-grounded. They can also manipulate the public into accepting illegitimate use of power by other authorities by appealing to their expertise (e.g., restricting human rights as "*necessary*" measures to deal with a pandemic). It is also hard for other types of authority, as well as for the public, to comprehend, evaluate, and criticise epistemic authorities due to their lack of expertise. Summing up, similarly to judicial authorities, epistemic authorities are unelected, unanswerable to the public, they can not be easily understood, evaluated, criticised, and they can use their power illegitimately against the interests of the public and its elected bodies. We saw in §2.4 that all those are more or less the reasons that judicial authorities have to be checked-and-balanced by providing *justifications* for their judgements that satisfy certain legitimacy requirements (e.g., establishing affordable procedures that allow for the public to access/request those justifications, the justifications being written in plain language without logical contradictions and ambiguity, and so forth). Consequently, the same justification requirements should be used to check-and-balance epistemic authorities (for a comparison between judicial and epistemic authorities *see also* Vibert 2007, pp.115-121).

AlgoAI engineers: a nascent epistemic authority

Any team of AlgoAI engineers constitutes an unelected epistemic authority. And *contra* to the usual bodies of such authorities like experts on food, environmental, or drug safety, AlgoAI engineers do not exercise solely epistemic power. Whenever the AlgoAI model they have engineered exercises any of the four powers, we have a *co-production*[52] of that power by both AlgoAI and its engineers. How does this co-production happens though? Many putative "*factual judgements*" about how an AlgoAI model should be engineered are in reality factual judgements in conjunction with evaluative (background) judgements similarly to the case of the Pareto optimality and the Rule of Law Index® that we saw in §1.2. More precisely, the AlgoAI engineers decide how to *interpret* values of a political order like racial equality or the right to life and how to *translate* those interpretations to components of AlgoAI models. Note that translating an interpretation of a value from a language $\mathcal{L}_1$ (e.g., ordinary language) to a language $\mathcal{L}_2$ (e.g., the formal language of first-order logic, a common language choice for legal AlgoAI (§3.2.1.1, ¶9; §II.4.2.1)) is an *interpretation* of values since the translators interpret expressions of $\mathcal{L}_2$ as having the same (or at least similar) meaning with expressions of $\mathcal{L}_1$. Since AlgoAI exercises power according to engineers' interpretations of the political order's ends, power is *co*-produced by *both* engineers and AlgoAI. It may be the case that AlgoAI has a certain level of autonomy when it exercises power like learning by itself how to exercise power (that autonomy is at the core of the threat of algocracy as we will see in §3.2.1.1). Still, even those (semi-)autonomous decisions are performed based on interpretations of concepts by the engineers like their interpretation of the concept of *learning*. Machine learning AlgoAI is bound to operate according to those interpretations (more on machine learning at *fn.* 81; *cf.* Danks 2014). Acknowledging the co-production of power is important for restricting the illegitimate use of power by the *engineers* of AlgoAI instead of focusing exclusively on the illegitimate use of power by the *users* of the said AI, a concern that is already present in the literature from the dawn of AlgoAI like in the case of the US military using AlgoAI systems during the Cold War: "*One danger of depending on elaborate simulations and computerized war games is that crucial decisions … tend to be made by the people who write the computer programs and build the elaborate model[…] The danger is rather that [the military commander] may depend on computerized decision aides without realizing how much human judgement has gone into making such aids useful to him.*" (Read 1961).[53]

replacement *v.* supportive AlgoAI

Let's see two examples of translation of values by legal AlgoAI engineers, one for each type of AlgoAI:

[52]The term "*co-produce*" that I am using is similar to Russo's 2022 concept of *co-production* of knowledge and ontology from both human and non-human agents (*ibid.*, §§9-10). Despite using some of Russo's arguments in this Thesis, I do not claim that my account of co-production is the same as hers. Note that Russo does not provide a full-fledged account of this co-production leaving it open-ended.

[53]I would like to thank dr. Stephanie Dick for introducing me to the early use of AlgoAI by the military & the subsequent legitimacy concerns during the Logic for the AI Spring 2022 summer school organised by Lake Como School of Advanced Studies as well as in personal communication. *See also* Keeny's 1986 "*Value-driven expert systems for decision support*", a paper published at a NATO 1896 conference on "*Expert judgment and expert systems*" (more on expert systems and their importance on legal AlgoAI on *fn.* 81; §3.2.1.1, ¶9; §II.4.2.1). As the title of the paper suggests, it is another effort to "*factualise*" evaluative judgements (*cf.* §1.2).

(α) *replacement* AlgoAI, i.e. AlgoAI that takes the role of a specific authority; (β) *supportive* AlgoAI, i.e., AlgoAI that constitutes a *means* that authorities use to exercise power (Winter, Hollman, and Manheim 2023, p.188). In the latter case, AlgoAI constitutes epistemic authority since it is used to process gathered information and output new information to support other authorities that exercise power. A frightening example of misuse of authority by experts when it comes to replacement AI is designing autonomous vehicles based on the *interpretation of the human right* to life by majorities *via* computational social choice tools (Etienne 2021). Apart from giving the ability to vote in favour of clearly unconstitutional choices (*ibid.*), this is a textbook case of institutionalisating mob rule (which is usually also unconstitutional) (*cf. fn.* 48). Regarding supportive AI, the most famous and controversial case due to allegations of racial profiling is COMPAS, an AI tool that estimates the probability of an offender recidivating (CEPEJ 2019a) based on an interview with the said offender and their criminal history (Winter, Hollman, and Manheim 2023, p.188). Similar tools have been HART and VICTOR used in the UK and Brasil respectively (*ibid.*). The controversy regarding COMPAS's *fairness* has raised issues regarding which formal translation of the value of fairness is the more adequate one (*see e.g.* Lagioia, Rovatti, and Sartor 2023, §3.7). For instance, a common formal translation of fairness is that of *counterfactual fairness*, according to which the normative is for the same algorithm to produce the same results given the same input regardless of certain protected characteristics like gender, race, etc. Imagine for example applying for a job, a parole, a loan, or a university, and an algorithm that reviews the applications exhibits a gender or racial bias among similar application like preferring male over female job applicants with the same qualifications.[54] A criticism of the counterfactual interpretation is that in social orders that are structured in ways that disadvantage certain groups of individuals (e.g., excluding women from education) it is *un*fair to give equal weight for the same qualifications both to those who had easier and to those that had more difficult access to those qualifications (Ali et al., n.d.; Cahoone 2023, pp.83-86).

Since AlgoAI engineers substitute judicial authorities in the interpretation of fundamental values, it is imperative for AlgoAI engineering teams to include experts outside of the AI discipline in order to make sure that such formal translations are indeed oriented towards the designated ends like legal and political scientists, philosophers of (meta-)ethics, and so on (henceforth *legitimacy experts*). In the CONCLUSION of this CHAPTER, I argue that AlgoAI engineering teams should also include *formal philosophers & logicians* as legitimacy experts (*cf.* §II.4.1.2). To further mitigate the distance between the engineers' and the judicial authorities' interpretations, the latter should also have their own *independent* teams of experts to aid them evaluating whether legal AlgoAI models abide by their interpretation of the law. If they do not, judicial authorities should strike them down as they do with unconstitutional legislation proposed by elected authorities. I.e., I propose an expansion of the scope of *judicial review* to include the fourth branch.

legitimacy experts

So far I have introduced checks and balances for bodies of experts that are *state*-actors. However, we saw that in contemporary political orders there are also *non*-state actors that exercise power like the media (§2.5, ¶1). Similarly to the case of the media, it is imperative for legitimacy to have *non*-state bodies of AlgoAI engineers that can *inform* the public about illegitimate exercise of power and that can propose policies to improve legitimacy (e.g., methods for producing justifications for legal AlgoAI's judgements which are sufficiently *understandable* by the public (*cf.* §III.3.2.4)). Such non-state-actors can be for instance non-governmental organisations (NGOs) emerging from civil society, also being called Civil Society Organisations (CSOs).[55] Note though that Vibert is skeptical with regards to bottom-up epistemic authorities: "*[top-down epistemic authorities] are also susceptible to manipulation by NGOs and so-called 'civil society organisations' that have a strong interest or advocacy position in the same field of activity.*", albeit they still acknowledge that NGOs can have a positive contribution to the bigger picture of the fourth power's legitimacy (*see e.g. ibid.*, p.163, footnote 34). Vibert's skepticism is justifiable since as authorities, bottom-up unelected can very well abuse/misuse power (e.g., being bribed to manipulate the public into accepting illegitimate AlgoAI). *Ergo* why it is necessary to acknowledge bottom-up unelected as part of the fourth power branch and subsequently apply the appropriate checks and balances.

bottom-up v. top-down epistemic authorities

The distinction between top-down and bottom-up epistemic authorities is not always clear. For instance, can non-state actors that receive funding or other types of support from state-authorities be truly independent from them? What about non-state actors that collaborate with state-atcors to aid them in exercising power like the CoE cooperating with NGOs for the implementation of the ECtHR's judgements?[56] Examples of (hybrid) bottom-up unelected relevant to the Thesis are WJP that publishes the Rule of Law Index®, EIU that publishes the Democracy Index, Oxford's Centre for the Governance of AI (GovAI) & Oxford Insights that

[54] Have a look at the Alan Turing Institute's *Counterfactual Fairness* research project and the paper of the same name published by its team of organisers: Kusner et al. 2017; *cf.* P. T. Kim 2022, §IV.B.

[55] "*A civil society organization (CSO) or non-governmental organizaiton (NGO) is any non-profit, voluntary citizens' group which is organized on a local, national or international level. Task-oriented and driven by people with a common interest, [CSOs]... bring citizens' concerns to Governments, monitor policies, and encourage political participation at the community level. CSOs provide analysis and expertise, serve as early warning mechanisms and help monitor and implement international agreements,...*" (UN's Civil Society Unit, accessed 02 June, 2023).

[56] *See* Rule 9 of the "*Rules of Procedure of the Committee of Ministers*"; Nussberger 2020, pp.159-160.

publish the Government AI Readiness Index (an index which provides an operational definition that measures the "*readiness to implement AI in the delivery of public services*" (Rogerson et al. 2022, p.6)), and the Irish Institute of International and European Affairs (IIEA) that publishes material about ALGOAI (*see* e.g. its paper on the relation between AI & the rule of law: Binchy 2022).

Regardless of whether they are top-down or bottom-up, if ALGOAI engineers are the philosophes of our time, and since we are again in a historical moment when a legitimacy paradigm is *disrupted* like what happened during Enlightenment, it is imperative for those new philosophes to transfer the discussion about this disruption from their sterilised isolated academic environment to the mainstream public sphere. They should act like philosophes both in their theoretical contributions and in their social activism ([*cf.* §2.3, ¶8]). Take the example of the British philosophe Thomas Paine, who made sure to publish and disseminate his seminal for the American Revolution "*Common sense*" (1776). While at the time both "*both "independence" and "republic" had become dirty words*", "*[in] just forty-six pages, Paine accomplished a remarkable feat—he showed that the monarchy was not divinely inspired but simply invented*". "*Common sense*" stripped "*the monarchy of its divinity [opening] the entire system up to critique, even mockery*" (Sunder 2020, p.999). As Sunder remarks, "*Common Sense* ***went viral****... As Eric Foner describes, between January and July 1776, "scarcely a week went by without a lengthy article in the Philadelphia press attacking or defending, or extending and refining Paine's ideas, and the same was true in other cities as well." [Foner 2005, p.74].*" (*ibid.*, emphasis added). Today, the new media of the cyberspace constitute such technological means with the Arab Spring being a prime example of this Enlightenment parallelism (Sunder 2020; *see also* the "*cyber-democracy*" entry in Campbell and Schneider 2020; Abu-Taieh, Hadid, and Zolait 2020). Waldron 2020 (§2) argues that the rule of law is a "*working*" political concept being shaped by "*ordinary citizens, lawyers, activists and politicians as of the jurists and philosophers*". The same holds for *legitimacy*: all three, the public, the authorities, and the philosophes ALGOAI engineers, *do* co-operate to a certain extent and they *should* co-operate even tighter to determine which aspects of the new legitimacy paradigm algocratic orders should accommodate. It is after all a *conceptual* necessity: since legitimacy is about an order being ordered according to the *true* beliefs of the subjects, it is imperative for the public to be *aware* of any changes in the political order's ends and *accept* them. Otherwise, their trust in that order is founded on *false* beliefs rendering it illegitimate.

socially re-weaponising reason

The legitimacy requirements laid out so far have been grounded on the *formal* conception of the rule of law. No matter how much one tries though, at one point, they will inevitably have to put *substance* in their requirements. In what follows, I delineate requirements about how ALGOAI engineering should be practiced so as to accommodate the different substances of the current world order according to the rule of law. To make my case, I introduce the concept of INDUSTRY 4.0, i.e. the technological advancements that made algocracy possible, and the concept of SOCIETY 5.0, i.e. the new paradigm of social order's structural dimension predicated upon the technological advancements of INDUSTRY 4.0. Those two concepts will allow me to bridge the theoretical groundwork I have laid out so far with the *actual* societal and political implications of the threat of algocracy. They will also provide me with further ammunition to explain & respond to that threat.

I.2.6 On the regional order of orders

We saw in §1.2, ¶4, that operational definitions can be employed to identify common legitimacy requirements across diverse legitimacy paradigms. We also saw that the formal non-substantive legitimacy requirements of the rule of law constitute such universal requirements. Such formal requirements though will not take us far. Each political order will have to add its own *substance* like democratic regimes adding requirements that realise the value of democracy (§2.8). Adding such requirements will result in checklists that look quite similar for political orders of similar legitimacy paradigms (henceforth *regions*). There is already a lot of literature in the discipline of *international relations* that groups the world into such regions like Kissinger 2014 or the "*prophetic*" Huntington (1996) 2011.[57] In what follows I argue why it is advisable for ALGO engineers to design models based on *regional* checklists *contra* state-specific or cross-regional international checklists. Regarding the differences between *world*, *international*, and *regional* orders, a *world order* is a *normative* concept held by a region about how the political order of the world should be ordered, an *international* order is the *practical application* of the normative concept of world order in more or less a global scale, and a *regional* order is its practical application but in a restricted regional area (e.g., in the European area) (Kissinger 2014, p.9). The arguments I am using to make my case are the *conceptual siilarity* across regional legitimacy paradigm (§2.6.1) and the practical conveniences in the Research & Development (R&D) & commercialisation of ALGOAI induced by engineering ALGOAI models based on regional checklists (§2.6.2).

adding *substance*

[57] To put things into perspective, while for a publication in international relations 1000 citations are considered a "*good*" number, by 2019, Huntington's book and his 1993 Foreign Affairs paper of the same name had 36 times that number (Haynes 2019, footnote 1). By June 2023, according to JSTOR, they reached 46.259 citations!

I.2.6.1 Why regional engineering: on SOCIETY 5.0

At the 2011 Hannover Fair,[58] the term "*INDUSTRY 4.0*" was coined to describe the technological developments of the German industry, developments that set it apart of the then ongoing 3rd Industrial Revolution (Schwab 2016, p.12; Philbeck and Davis 2018, p.17). 5 years later, the founder and executive chairman of World Economic Forum (WEF) Klaus Schwab argued that what was happening to Germany was in reality a global scale phenomenon, the phenomenon of the *4th Industrial Revolution* (or *4IR*). Schwab popularised the "*4th Industrial Revolution*" term through the publication of his 2016 book of the same name and by chairing the 2016 WEF of the same official theme (Wearden 2016). After that, INDUSTRY 4.0 became a synonym with the 4IR (*see e.g.*, McKinsey *&* Company 2022; Mourtzis, Angelopoulos, and Panopoulos 2022; Bai et al. 2020, §Abstract), with previous industrial revolutions being named similarly as INDUSTRIES 1, 2, *&* 3. In the midst of all that, already from 2015, the Japanese government coined the term SOCIETY 5.0, the transformation of society based on the technological advances of the 5th Industrial Revolution (or INDUSTRY 5.0) (Mourtzis, Angelopoulos, and Panopoulos 2022, p.7),[59] as a response to the European-centered INDUSTRY 4.0 and China's MADE IN CHINA 2025 socio-technological plan (*ibid.*, p.3). The velocity of those changes should not come as a surprise since already from 2016 Schwab was acknowledging that *contra* to the previous industrial revolutions, INDUSTRY 4.0 was progressing *exponentially* rather than linearly (Schwab 2016, p.8).

The threat of algocracy is a threat that concerns the legitimacy of SOCIETY 5.0's political orders. It is certain characteristics of SOCIETY 5.0 that necessitate ALGO engineering to be conducted at a regional level. And those specific characteristics became possible due to INDUSTRY 4.0's technological advancements. Consequently, to make my arguments, I will first introduce in more detail the concepts of SOCIETY 5.0 and INDUSTRY 4.0 starting from the latter.

A revolution is the result of a *disruption* of actuality, like the emergence of new technologies disrupting LAW 1.0 and leading to LAW 2.0 or the use of reason as the ultimate interpreter of the world disrupting the pre-Enlightenment legitimacy paradigm of faith. Similarly, INDUSTRY 4.0 spawned after the disruption of INDUSTRY 3.0 (Schwab 2023). Specifically, INDUSTRY 1.0 exploited "*steam power to mechanize production*", INDUSTRY 2.0 "*was driven by mass production made possible through electricity*"', and INDUSTRY 3.0 that began at the 60's introduced "*digital technology to automate production*" (Park 2016, p.1; *see also* Philbeck and Davis 2018, pp.18-19). INDUSTRY 4.0 is an *epi*-digital revolution that *disrupted* INDUSTRY 3.0's separation of the digital and physical world merging them to novel *cyber-physical* spaces. Further disruptions are the connectivity among INDUSTRY 3.0 infrastructures and the ability to harvest big data with high computational power (e.g., cloud technology, the Internet of Things (IoT), blockchain, quantum computers), advanced AI (e.g., advanced leaning skills and higher autonomy compared to older AI (more on §3.2.1.1 *&* *fn.*79)) and advanced engineering (e.g., nanotechnology, genome editing, 3-D printing, virtual reality (VR)) (McKinsey *&* Company 2022; WEF 2019, p.8; Philbeck and Davis 2018, pp.20-21). As we will see later on, the technological advancements that are of central importance to the threat of algocracy are the emergence of the *cyber-physical space* and the advancements in *self-learning*, *decision-making*, and *pattern recognition* in contemporary AI. What about SOCIETY 5.0 though? How did SOCIETY 5.0 emerge through INDUSTRY 4.0's disruptions?

INDUSTRY 4.0

SOCIETY 1.0 was the human-gatherer society, SOCIETY 2.0 the agrarian society, SOCIETY 3.0 the industrial society that emerged during INDUSTRY 1.0, and SOCIETY 4.0 is the society that emerged during INDUSTRY 3.0 until today. It is the *information* society, that became possible *via* the advancement of Information and Communication Technologies (ICTs)[60] and the introduction of *cyberspace*.[61] In SOCIETY 4.0, the cyber and the physical

SOCIETY 5.0

[58] As its German name "*Hannover Messe*" suggests, it is a trade fair, one of the most, if not *the* most, important in the world: https://www.hannovermesse.de/en/ (accessed 10 June, 2023).

[59] *See* Japan's Cabinet Office's https://www8.cao.go.jp/cstp/english/society5_0/index.html and https://www.japan.go.jp/abenomics/_userdata/abenomics/pdf/society_5.0.pdf (accessed June 10, 2023). As we will see multiple times later on, Japan has a history of implementing highly ambitious AI policies toward the re-engineering of its social order. In 1981, it put forward the first national AI-oriented project in the world, the so-called FIFTH GENERATION COMPUTER SYSTEMS (FGCS) project (Nitta and Satoh 2020, p.473). It was a 10years $1.3+ billion plan to build "*massively parallel, intelligent*" computers using the logical programming language PROLOG (Russell et al. 2021, p.41), a project that spawned a spring of *legal* logic-based AI in Japan (Nitta and Satoh 2020, p.471). In 2004-2010, Japan put forward the e-SOCIETY project, a precursor of SOCIETY 5.0 whose goal was to digitalise a big part of social activity including laws and social customs using once more logic-based AI (§II.4.2.1). Additionally, in 2017 till today, it put forward the ADVANCED REASONING SUPPORT FOR JUDICIAL JUDGMENT BY ARTIFICIAL INTELLIGENCE project which is pretty much self-explanatory (§II.4.2.3).

[60] Examples of ICT are navigation/video/audio/data networking equipment, broadcasting and communications services, personal computers, supercomputers, software, mobile and cloud platforms (Byrne and Corrado 2016, SS2-4). Based on the foregoing, ALGOAI belongs is ICT.

[61] The term "*cyberspace*" was firstly coined by the cyberpunk author Ford Gibson in his 1984 "*Neuromancer*". Minimally, cyberspace can be used as a synonym to the *Internet* (Gálik and Tolnaiová 2020, p.13). Such a restrictive definition is obsolete in the context of SOCIETY 5.0 in which I contextualise the threat of algocracy. Whenever I clean my room together with my cobot (collaborative robot), we interact in a cyber-physical space which is *not per se* connected to the Internet. I adopt Gálik and Tolnaiová's non-minimal construal of cyberspace as any non-physical space generated by ICT, from the old-school telegraph (Gálik and Tolnaiová 2020, pp.13-17) to social media platforms.

space, the digital and the analogue *infosphere*,[62] remain more or less distinct. It is INDUSTRY 4.0 that brought them together, setting the foundations for a new cyber-physical society, *SOCIETY 5.0*. However, INDUSTRY 4.0 was about introducing disruptive technology in the *industry* sector, technology useful to the experts of specialised industrial tasks. Here is where INDUSTRY 5.0 comes to the rescue: INDUSTRY 5.0 *disrupts* the industry-centered character of INDUSTRY 4.0 making it a *human*-centric one so as to solve *social problems* (Mourtzis, Angelopoulos, and Panopoulos 2022, §2.4; Carayannis and Morawska-Jancelewicz 2022, p.3448; M. Fukuyama 2018, p.48). It is about the merging of the cyber-physical with everyday social activity allowing the relevant technology (like big-data AI) to be used by ordinary citizens and not domain experts so as to co-produce knowledge and ontology that resolves their problems (*cf. fn.* 52). For instance, Japan aims at using the advancements of SOCIETY 5.0 to meet UN's 2030 Sustainable Development Goals (SDGs)[63] like promoting peace, justice, and strong institutions, reducing inequalities, and ensuring healthy lives and well-being for all at all ages (Bai et al. 2020, table 2; Folarin, Akinlabi, and Atayero 2022). Note that some reject the position that INDUSTRIES 4.0 & 5.0 are distinct; there is *only* INDUSTRY 4.0.[64] For instance, Bai et al. 2020 not only argue that INDUSTRY 4.0 suffices to realise the human-centric SDGs, but they even introduce a new mathematical measure (i.e., a "*factualised*" operational definition) to evaluate the impact that different technologies of INDUSTRY 4.0 have to their realisation. Even if INDUSTRY 4.0 & 5.0 are not distinct, SOCIETY 5.0 is clearly not the same as SOCIETY 4.0 since in the former, the physical and the cyber space are distinct. Concluding, the core distinctive characteristic of INDUSTRY 5.0's technology is that algorithmic models "*perform or support [...] the work and adjustments that humans have done up to now*" (Carayannis and Morawska-Jancelewicz 2022, p.3449). Therefore, ALGOAI is a quintessential example of INDUSTRY 5.0 technology which is already used to transition to SOCIETY 5.0.

For SOCIETY 5.0 to be legitimate, both the *physical* and the *cyber* space need to be well-ordered. However, the fact that cyberspace eradicates physical borders renders *state*-authorities powerless to well-order it just by themselves. Together with *global change*,[65] cyberscpace's *transnational*[66] unorderness constitute typical arguments in favour of allowing political actors whose authority transverses national borders to trump national sovereignty and exercise *supranational* power (e.g., what the ECtHR does).[67] When an advertisement company in the Netherlands data profiles internet users in the Balkans using AI-assisted software developed in Israel, if Dutch and Israeli authorities do not cooperate, the Balkan state-authorities can not place legally-binding restrictions in the use and development of AI neither in the Netherlands nor in Israel. Child pornography (Serebrin 2023; Ratner 2021), cyberterrorism and cyberwarfare (Baker-Beall and Mott 2021; Kissinger, Schmidt, and Huttenlocher 2021, pp.139-141; Dinniss 2018; Christina 2017 *contra* Jacobsen 2022), phishing (Fritsch, Jaber, and Yazidi 2022), hate speech (Vidgen, Burden, and Mergetts 2021; Bayer and Bárd 2020), spyware and cyberviolence like revenge porn (*see* §2.1. ¶7) are few of the most common *transnational* problems that require *supranational* responses. Why though should that supranational response be *regional* and not international?

supranational solutions for *transnational* problems

In the previous paragraph, I made an analogy between cyberspace's transnational unorderness and global change. That was no accident. Cybespace is still *space*, a non-physical space, but still a space. And by being *non*-physical, cyberspace is by default untamed by physical national borders, just like nature.[68] However, *contra* to nature being shaped by *natural* laws that hold all around the globe, different social orders are shaped by different laws, different from legitimacy paradigm to legitimacy paradigm. Consequently, *regions* of the world with similar legitimacy requirements can cooperate to establish their own *regional* cyberorders making sure that

the *regionality* of law *v.* the *internationality* natural law

[62] "***Minimally**, infosphere denotes the whole informational environment constituted by all informational entities, their properties, interactions, processes, and mutual relations. It is an environment comparable to, but different from, cyberspace, which is only one of its sub-regions, as it were, since the infosphere also includes offline and analogue spaces of information [e.g., libraries, printed books and newspapers, museums]. **Maximally**, infosphere is a concept that can also be used as synonymous with reality, once we interpret the latter informational.*" (Floridi 2014, p.41). I adopt the minimal conception of the infosphere in order to separate between analogue and digital sources of information that correspond to pre- and post-cyberspace social orders respectively.

[63] *See* the first URL in *fn.* 59.

[64] "*At this point, it has to be stressed that I4.0 is an ongoing technological evolution, and Society 5.0 (including Industry 5.0) is still under preparation, thus creating a misconception that Industry 5.0 will not be considered as an independent industrial revolution.*" (Mourtzis, Angelopoulos, and Panopoulos 2022, §1).

[65] "*Global change*" refers to the "*planetary-scale changes [that] are occurring rapidly*" and are caused mainly from human activity (Steffen et al. 2005, p.4). It does not include only climate change, but also all other environmental changes like exhaustion of natural recourse (wood, petroleum, etc) and species extinction, as well as the implications of those changes on human society (economy, living standards, and so on) (*see ibid.* pp.3-9; Pranab, Nandan, and Kalyan 2017, pp.1-3).

[66] *Transnationalism* "*denotes the social and global transformations of interconnectivity between peoples, states, economies, and cultures under the processes of globalization*" (Brown, McLean, and McMillan 2018).

[67] *See e.g.* Weart 2023; Kikarea and Menashe 2019; *cf.* Raymond 2013; *contra* Kikarea and Menashe 2019. "*Supranationalism*" "*[r]efers to the formal transfer of legal authority and decision-making power from member states to an institution or international body.*" (Brown, McLean, and McMillan 2018). The ECtHR (Scheeck 2005) and the EU (Cafaro 2023, pp.66-69) are such supranational bodies.

[68] The inability of a single state to deal with environmental changes can be seen by the fact that the very few cases where the applicants requested from the ECtHR to protect their right to life in the face of climate change are cases against not one or two or three HCPs, but *tens* (!) of them. E.g., *Duarte Agostinho and others v. Portugal **and 32 other states**, De Conto v. Italy and **32 other states**, Soubeste and four other applications v. Austria and **11 other states**, Uricchiov v. Italy and **31 other states***. More on ECtHR's Press Unit 2023a.

the respective laws are oriented towards their common ends (for a similar argument *see* Raymond 2013). At the same time, there should be *cross*-regional cooperation to find common ground on transnational problems shared between the regions. However, since any such compromises will have to be compatible with all involved regional legitimacy paradigms, they will not be able to realise adequately the *substance* of the different legitimacy paradigms. Consequently, due to the *conceptual similarity* among legitimacy paradigms, AlgoAI engineers should first prioritise regional legitimacy requirements among orders with similar substance, and then perform further adjustments to fit international and state-specific requirements.

Finally, it should be noted that the digitalisation of political activity in society 5.0's cyber-physical space results in more and more *non*-state-actors exercising political power in the place of state-authorities like social media companies regulating the human right to freedom of expression (article 10) (*cf.* Heldt 2019; Barrett 2020; Sartor and Loreggia 2020, p.29). Rules imposed by non-state-actors in the cyberspace were labelled by the UN as "*platform law*" (UN A/HRC/38/35, ¶1) like "*Facebook law*" and "*Twitter law*" which are "'*displacing the laws of national jurisdictions*" (Land 2020, p.975). Such platform laws tend to be in line with the opinions of the majority (i.e., the users of those platforms) *contra* the anti-majoritarian character of human rights (Robertson 2004, pp.49-50; *cf.* §2.8, ¶2) leaving vulnerable groups unprotected. And all that in conjunction with an *opaque* application of those rules.[69] At the same time though, making rules based on what the majority believes corroborates the legitimacy value of democracy (*ibid.*, pp.975-976; *cf.* §2.4, ¶8; §2.8, ¶2). Considering these, AlgoAI engineers need to make sure that AI engineered to apply platform law (e.g., *upload filters*[70]) is engineered in accordance with the measures taken by state-authorities to regulate those non-state-authorities: "*The United Nations, regional organizations and treaty bodies have affirmed that offline rights apply equally online, but it is not always clear that the companies protect the rights of their users or that States give companies legal incentives to do so.*" (UN A/HRC/38/35, ¶1).

regulating *non*-state-actors' authority

I.2.6.2 Why regional engineering: on the pragmatic effects of legitimacy

Choosing to develop AlgoAI based on regional legitimacy requirements leads to a handful of practical advantages from less cost on Research & Development (R&D) (e.g., the same technology can be used by multiple political orders) to the adoption of *common strategies* to deal with legitimacy-induced transnational challenges in the *R&D* and *commercialisation* of AlgoAI. Once more, transnational problems urge for regional supranational collaboration, but this time not for reasons of *conceptual similarity*, but of *practical convenience*. Let's see in more detail why this is the case.

On the one hand, *prima facie*, it seems that legitimacy paradigms that protect human rights and democracy are advantageous to Research & Development (R&D). In an award-winning article for the Journal of Law in the Middle East, Araujo 2022 argues that the fear of severe legal repercussions induced by the Middle Eastern Shari'a-based legal system can potentially discourage innovation in AI engineering pratice in the Middle East region *contra* the liberal Western regions. Eminent philosopher Karl Popper made similar arguments (Popper 2005, 2012) contending that individual and economic freedom fostered by liberal democratic states, as well as the protection of property rights, can advance innovation. Arguments that have been recently backed by empirical evidence (Wang et al. 2021).

Enlightenment values *fostering* R&D

Reality though is far from being that black-and-white. Protecting human rights and democracy entails *restrictions* on the development and integration of Algo technology with unfavourable implications. In a statement before the US Senate Judiciary Committee, Layton 2019 (experts on international technology policy) warned that EU's benchmark for data-protection *General Data Protection Regulation* (*GDPR*; *cf.* §3.2, ¶2) is costly, weakens small and medium enterprises (SMEs), hinders academic research, and eventually strengthens the largest players. At the same time, while China and the EU share similar ethical and legal concerns about the use of AI, China's Confucianism-based legitimacy values and Europe's Enlightenment-based ones lead to very different responses to those concerns that give the former a lead in the AI race. Specifically, China follows a *promotional* approach based on the public's trust that the government will be dealing with legal/ethical concerns as it moves towards the realisation of future goals (e.g., since 2017 China has in place the "*New Generation Artificial Intelligence Development Plan*" ("新一代人工智能发展规划") strategy aiming at becoming the world AI leader by 2030 and to monetise AI into a 150 billion dollars industry (Roberts et al. 2021)). Chinese citisens are more willing to compromise with temporary suboptmial resolutions of ethical/legal concerns as long as there is progress towards the desired ends. Europe on the other hand follows a *prohibitive* approach imposing limitations on the development and integration of AlgoAI slowing down its development. It is an approach that stems from the traditional Enlightenment-rooted liberal distrust of government (Fung and Etienne 2022).

Enlightenment values *hindering* R&D

[69] More on the *opacity* of AI on §3.2.

[70] *Upload filter* are ADMs that filter the content uploaded by internet users like censoring speech that they classify as hateful (Sartor and Loreggia 2020, pp.9-10,41-42; *cf.* Heldt 2019).

Now that I have made my case of why ALGOAI engineering should be performed *regionally* so as to accommodate the substance of different political orders, it is time to introduce a few examples of that substance for the case of the European political order. That substance concerns the values of *human rights & democracy* and their relation to the *rule of law*. Regarding *human rights*, I still choose to provide *minimal* substantive requirements that are shared among most regional legitimacy paradigms. Ergo, one could claim that to some extent they are still *formal* requirements. Regarding *democracy*, I highlight two topical conflicts about its substance, with one of the two being the main premiss in favour of abandoning Enlightenment's legitimacy paradigm.

I.2.7 On human rights

I begin this subsection by introducing what human rights *are & why* they are of fundamental importance for legitimacy. I conclude by extrapolating two legitimacy requirements, the so-called *minimality* requirements.

The fundamental intuition of the post-1945 concept of *human rights* is that they are rights one possesses under the sole condition of *being human* (Holm 2023; Nickel 2021), *contra* other rights, like the right to get exempted from municipal taxes (unfortunately). As expressed by their Enlightenment precursors (Weston 2023; Nickel 2021, §3.1,§2.1; Fagan, n.d.):

> "*Article 1: **Les hommes naissent** et demeurent libres et égaux en droits*"
>
> *Déclaration des droits de l'homme et du citoyen* (30 September 1789),[71] emphasis added

> "*...**all Men are created** equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty,...*".
>
> *The Declaration of Independence* (4 July 1776), emphasis added

That being said, human rights are not mutually exclusive with other types of rights like civil rights (e.g., the Convention protects the civil right to fair hearing by an independent and impartial tribunal (ARTICLE 6, ¶1)) or environmental rights (e.g., in its case-law the ECtHR has judged multiple times that environmental changes have violated the RIGHT TO LIFE (ARTICLE 2) like *Brincat and others v. Malta, 2014*; *L.C.B. v. the UK, 1998*; *Öneryıldız v. Turkey, 2004*; *Budayeva and others v. Russia, 2008*; *Mučibabić v. Serbia, 2016*; *cf. fn.* 68) (Nickel 2021). Probably the biggest controversy about which types of rights are human rights is about whether *labour rights* are human rights (*see e.g.* Mantouvalou 2013 and ECtHR's Press Unit 2023d). The human rights protected by the Convention that are of particular importance for this Thesis are the rights to a *fair trial & due process* (ARTICLE 6), *privacy and data protection* (ARTICLE 8), *freedom of thought, conscience and religion* (ARTICLE 9), *freedom of expression and information* (ARTICLE 10), *enjoy human rights and freedoms without any discrimination* (ARTICLE 14). For a concise summary of potential threats to those rights by AI *see* MSI-AUT 2019, §2.1; Commissioner for Human Rights 2019; the "*European ethical Charter on the use of artificial intelligence in judicial systems and their environment*" (CEPEJ 2018). All of those documents are prepared by top-down unelected CoE authorities.

Why though are human rights that important for legitimacy? The intuition that human rights are rights one has just by being human is also the core intuition behind the property of human rights *being universal*, a property to which everyone seems to agree albeit with substantial disagreements about its meaning (Nickel 2021, §1).[72] More precisely, the "*[o]rthodox [v]iew*" of human rights is that they are both *universal* and *moral* (Tasioulas 2011, p.20; *cf.* Lefkowitz 2020, pp.131-133), where *being moral* entails that the subjects do not have a mere *legal* obligation to obey them, but a substantive *moral* one (remember the distinction between legal and ethical concepts in §1.1, ¶2) (Tasioulas 2011, pp.27-28; Stanton-Ife 2022). The distinction between moral and legal obligations, a central if not *the* central question in jurisprudence, is of key importance for legitimacy. As argued in §2.2, ¶10, subjects obey rules if they believe that the authorities issuing the rules satisfy certain *normative* ends. If a subject has reasons to believe that those ends are *moral*, then their belief that this rule should be obeyed becomes stronger. At the same time and more importantly, if the subject believes that a rule is *im*moral, then it is more likely that they will believe that the rule should not be obeyed (Stanton-Ife 2022) and subsequently the authority that issues such immoral rules is *de*legitimised. Considering this, in case of a conflict between a legal and an ethical legitimacy requirement, if the engineers choose to prioritise the legal one as they should due to the value of legality (§2.4, ¶3), then the rejection of the ethical norm will come with a cost to legitimacy's *epistemic* dimension. This is an example where an appropriate public *justification* of the engineers' rejection of the ethical requirements could reduce the harm to legitimacy. The necessity for such justifications as well as the attempts to satisfy both legal & ethical requirements is also another example of why experts from other disciplines like law & philosophy should join ALGOAI engineering teams.

ethical *v.* legal requirements

[71] https://www.elysee.fr/la-presidence/la-declaration-des-droits-de-l-homme-et-du-citoyen (accessed 01 May, 2023).

[72] Remember e.g. §1.1, ¶3, where we saw the debate that the ECtHR judges had about the operational definition of human rights'

The human rights legitimacy requirement that I want to introduce is that of *minimality* due to its gravity and its almost universal character. *Minimality* can be construed in at least two dimensions. The first one is that human rights have *priority* over other rights; they are the *bear minimum* of rights that a political order ought to protect. They are "*...the least that every person can demand and the least that every person, every government, and every corporation must be made to do.*" (Shue 1980, p.ix; *see also* Nickel 2021, §1; CDL-AD(2016)007, ¶16). Certain human rights entail further obligations for judicial authorities to provide *justifications* for their judgement. For instance, legal AlgoAI that is used in criminal or civil law trials in the European order should provide a justification for its output to the involved parties in accordance with the justification requirements imposed by ECtHR's case-law on ARTICLE 6 (RIGHT TO FAIR TRIAL) (ECtHR's Registry 2022b, pp.38-41; 2022, pp.96-97; MSI-AUT 2019, pp.29-30; *compare to* EU's GDPR, Article 22). I will call such rights as *rights to justification*.[73]

Human rights: the *bear* minimum

The second dimension of minimality is that despite being a priority human rights should not be "*too demanding*"(J. Nickel 2021, §1). More precisely, different political orders should have the *margin* to decide by themselves how to realise them. In the ECtHR legal tradition, this is expressed *via* the so-called *margin of appreciation*. According to Letsas' seminal paper on the topic, in the ECtHR case-law, there are two different concepts of *margin of appreciation*: a *substantive* and a *structural* one. According to the *structural* concept, the ECtHR has certain restrictions in its authority to trump national legislation therefore respecting the sovereignty of the HCPs (*ibid.* pp.720-721). It further entails that the AlgoAI engineers should not *overdo* it by requiring the model to do more than what the state-authorities have decided. If they do so, then they impose their *own* interpretation of human rights in the model; it is the engineers that *appreciate* what measures should be taken and what restriction on human rights should be imposed. Unless the right to exercise such power is legislated (LAW 2.0), this constitutes a violation of the principle of legality (CDL-AD(2016)007, p.11). The structural conception of the margin of appreciation further entails that any AlgoAI model of the European area should abide by the case-law of the ECtHR that holds *for every* HCP and *only* for that case-law. Now regarding the *substantive* concept of the margin of appreciation, it requires that the HCPs should take *extra* measures than those imposed by the ECtHR, but they have a margin to appreciate *which* measures to take to protect human rights as well as when human rights shall be *restricted* (Letsas 2006, pp.709-710). The substantive conception entails that in each HCP, AlgoAI should be adjusted to the human rights law of that particular state order.[74]

Human rights: *not* too demanding

As argued in the §INTRODUCTION, in the post-WWII international order, human rights at least in a minimal form have been accepted as *normative* ends for more or less every state order, albeit many times this is a mere facade. Based on this universality, as well as on the minimality requirements, on could claim that human rights, at least a subset of them, are borderline *formal* and *not* substantive requirements. In the next subsection, the difference between substance & formalism is anything but fuzzy.

I.2.8 On democracy & epistocracy

> *"The world's only superpower is rhetorically and militarily promoting a political system that remains undefined–and it is staking its credibility and treasure on that pursuit."*
>
> *The struggle for democracy*
> Horowitz, 2006

As argued multiple times so far, the meanings of the legitimacy values are highly contestable. If there is a value that takes the lead in controversy, that is *democracy* (EIU 2023, pp.64-66). Two controversies are of importance for this Thesis. The first is about the topical *liberal v. illiberal* democracy debate, a debate that exhibits in the best way the differences between *substantial* and *formal* legitimacy requirement. The second controversy is about the *instrumental v. non-instrumental* democracy debate, a controversy which is in the epicenter of the threat of algocracy even for *non*-democratic political orders.

The *liberal v. illiberal* democracy debate is a debate about the *substance* of a political order. The term "*illiberal democracy*" is attributed to Fareed Zakaria's 1997 "*The Rise of Illiberal Democracy*" Foreign Affairs paper (Plattner 2019, pp.7-8). Illiberal democracy refers to democratic orders where majoritarian values are generally prioritised over counter-majoritarian ones. For instance, judicial review is restricted resulting in elected authorities acting more independently and individual rights (including human rights) being restricted when they conflict with the views of the majority. Hungary is a quintessential paradigm of contemporary illiberal democracies with its PM Victor Orbán consistently undermining the rule of law (*ibid.*; *cf.* §INTRODUCTION). The substance of illiberal democracy is reflected in the following 2014 Orbán quote: "*Hungarian nation is not a*

liberal *v.* illiberal democracy

universality.

[73] This term is motivated by Goodman and Flaxman's 2017 *right to* ***explanation*** used to describe EU's GDPR justification obligations. For the difference between *explanation* and *justification see* §IV.2, ¶3.

[74] For a "*rethinking*" (sic) of the two concepts of margin of appreciation *see* Arnardóttir 2016.

simple ***sum of individuals****, but a* ***community*** *that needs to be organized, strengthened and developed... the new state that we are building is an* ***illiberal state****... It* ***does not deny*** *foundational values of liberalism, as freedom, etc. But it does* ***not*** *make this ideology a* ***central*** *element of state organization, but applies a specific, national, particular approach in its stead.*" (Cahoone 2023, p.83, emphasis added). In other words, individual rights are not abolished, but they are restricted in the name of the "*community*". These substantive characteristics of illiberal democracies are why epistemic accessibility requirements like legal certainty, foreseeability, and open government are *weak* formal requirements (§2.4, ¶8). The law should still be epistemically accessible by the public, but not *per se* from everyone and not *per se* on the same degree. It suffices for the privileged majority to know which are the laws. Minorities having less epistemic access to the law contributes to the restriction of their rights. At the same time, the lack of foreseeability makes minorities more susceptible to acting against the law and hence providing further justification for the restriction of their rights. It is a vicious circle.

The *instrumentalist v. non-instrumentalist* debate is about whether democracy should be an *end* of a political order's functional dimension or whether it should be construed as *means* to realise other ends (e.g., human rights or economic prosperity). Instrumentalists that support democracy like John Stuart Mill (*see e.g.* Mill 1901) do so by arguing that democracy is more adequate than its alternatives to realise those ends. A consequence of this position is that if we are able to identify non-democratic means that realise more adequately those ends, like unelected AlgoAI actors drafting and enforcing legislation, we should not be hesitant to adopt them. A *non*-instrumentalist would reject such means on th ebasis that there are certain democratic values that are ends in themselves. Such a value is *political equality* (Peter 2017, §§3.2,4.1), where political equality can be construed as the position that "*all citizens should be treated equally in the democratic process*" like everyone having the right to vote (Blau 2023, p.23). John Rawls (1921-2002), seminal philosopher of law & politics as well as advocate of the non-instrumentalist position, argues that it would be "***irrational***" to "*give up the right to vote, even if that would massively improve their welfare, because this would be "humiliating," "destructive of self-esteem," and would express the idea that they are* ***subordinate***" (Brennan 2016, p.125, emphasis added). Rawls' argument pre-supposes *inter alia* that the value of *self-governance* (or *personal autonomy*), another Enlightenment legitimacy value (§2.3, ¶8) with the universal adult suffrage being a typical realisation of that value, should be an *end* in itself. As we will see, self-governance is of central importance for the threat of algocracy. A minimal construal of a *self-governed actor* is that of an actor that has the *authority* to decide by themselves how they will act (Buss and Westlund 2018, §3). Note that even if one abolishes their right to self-governance under an instrumentalist premiss, as long as this decision is justified *via human reason*, we are *still* inside the Enlightenment's *utilitarian* legitimacy paradigm (Peter 2017, §§3.2,4.1). It is not that easy to kill Enlightenment after all.

instrumentalist *v.* non-instrumentalist democracy

An alternative to democracy that constitutes the foundation of the threat of algocracy and that is grounded on the instrumentalist premiss of optimising a political order's ends is *epistocracy*. "*[I]f we assume (plausibly) that* ***legitimacy-conferring*** *outcomes are more likely to be achieved by those with better epistemic abilities*", then those should govern us (Danaher 2016, p.250). It is a restatement of Plato's *Philosopher King* position from his very democratic "*Republic*"[75] which is summed up to the following *dictum* (henceforth the *epistocratic principle*): "*[T]hose who have a special epistemic position (the wise, the educated, the knowledgeable) should rule.*" (Kuljanin 2019, p.81). Kuljanin 2019 provides a concise summation of the *epistocratic argument* (p.82):

epistocracy

ONTIC TENET: there exist correct procedure-independent answers to (some) political questions;

EPISTEMIC TENET: some actors are more likely to identify those answers. Those are epistemically privileged;

AUTHORITY TENET: the epistemically privileged should have political authority *in virtue of* their epistemic privilege.

ANTI-AUTHORITY TENET: the epistemically *under*privileged should *not* have political authority *due to* their epistemic disadvantage.

By "*procedure-independent*", one means that there is nothing inherent "*in democratic procedure that makes it very likely to come up with correct answers*" (*ibid.*, p.81). In the context of this Thesis, by "*epistemic underpriviledged*", I construe those that can not make *rationally* justified decisions as frequently as necessary for a legitimate exercise of power. E.g., while it is inevitable for human judges to make a certain number of biased judgements, that number should remain as low as possible for their authority to be legitimate (Chatziathanasiou 2022, p.455; *cf.* §3.1, ¶1; §II.3.1.2.2). If that number is not lowered below an acceptable level, then those judges lack the

[75] Another worth-mentioning revival of the Philosopher King position is that of utilitarian and instrumentalist supporter of democracy John Stuart Mill who argued that "*political rights should be (nearly) universal, but not equal – educated and professionals should have more votes than uneducated or menial labourers.*" (Kuljanin 2019, p.81).

required epistemic abilities and ergo they are epistemically underpriviledged. The authority tenet is attributed to Estlund 2008 (*cf.* Estlund 1993; 2003) and it is the one that Danaher 2016 uses in his "*The threat of algocracy*" paper. The anti-authority tenet is credited to Brennan's 2016 "*Against democracy*". Despite Danaher 2016 not using the anti-authority, it is still of high relevance to the threat of algocracy especially for the threat induced by *replacement* AlgoAI.

The epistocratic argument does not apply only to *elected* authorities. Bodies of *un*elected like epistemic and judicial authorities consist of individuals who are expected to be the most adequate ones to *epistemically* access an outcome that is the closest possible to the correct procedure-independent answer. Remember for instance that the rule of law imposes the obligation for the judges to be "***competent** enough to deliver just judgements, accompanied with adequate justifications written in plain language and in a timely manner*" (§2.4, ¶7). Furthermore and more importantly, even judges have to *vote* when making a judgement,[76] and ergo, for the epistocrats, the judges' voting rights should be regulated based on the (anti-)authority tenet. Why should every judge have the right to vote or the right to an equal vote if some of them (in collaboration) are more likely to identify the optimal solution? What if the others hold them back from reaching that ideal? These are more or less the basic epistocratic premisses that ground the abandonment of the Enlightenment legitimacy paradigm in contemporary algocracies. But first things first, *what is algocracy?*

epistocraty *v.* judicial *&* epistemic authorities

I.3 On algocracy

In the Introduction, I introduced algocracy as a "*governance system that uses ADMs to exercise power*". Let's restate this vague definition using the conceptual framework of this chapter: *algocracy is a political order where political power is exercised inter alia **by** or **via** ADMs*. I wrote "*by or via*" since algocracy concerns both *replacement* and *supportive* AI (§2.5, ¶9). I also wrote "*inter alia*" to state clearly that algocracy does not have to be a cyber-dystopian political order where "*artificial agents seize control of governmental decision-making bodies and then exercise power in way that serves their needs and interests*" (p.247). We do not have to wait for a Skynet to take over. *All* contemporary states are more or less algocratic; we are *already* confronted by both algocracy's perils and perks. And I am saying "*perks*" since neither Danaher nor I use the suffix "*-cracy*" pejoratively like many do in the cases of "*bureau-cracy*" or "*techno-cracy*" (*ibid.*). A real-life example of how the same (AI-assisted) algocratic technology can be used to both realise and undermine the same Enlightenment-based ends is the post-Arab Spring MENA's political order (Švedkauskas 2022, pp.39-40; Kausch 2022, pp.84-86).

What is algocracy?

I.3.1 On the perks

In what follows, I summarise the main positive impacts that current AlgoAI has in the legitimacy of algocratic political orders. Note though that I do not *per se* agree with every argument used to support those advantages in the cited sources. For instance, in order to support that AI can apply the law with less unbiased that human judges, Korean citisen and data analyst Yeonsoo Doh (2020), argues that judges should not be more lenient when the defendant of a sex crime submits an "*apology letter*". Although I disagree with that specific argument, a genuine apology, as well as an ingenuine apology or an unremorseful lack of apology, should be taken into consideration by the judges, the alleged reduction of bias is one of the most characteristic advantages of legal AlgoAI and it can be supported without that argument.

> "*Unlike humans, AI judges will not be swayed by neither personal connections, sentiment, nor bribe. AI judges will not accommodate offenders, because they will not have any personal connections. They will not reduce the sentence even if the offenders were drunk or the offenders submit letter of apology to the judges. They will not be bribed because it is no use. Yet AI judges will only stick to the code of laws and judicial precedent. Thus, they will make decisions logically once through deep learning of good judicial precedent.*"
>
> *Why the South Korean public wants AI judges*
> Doh, 2020

These are some of the arguments that raise the Korean public's support of replacing judges with AI models, with many submitting petitions to the presidential office (the so-called Blue House) officially requesting such replacements (*ibid.*; Ah-hyun 2021). The gist of the argument is that AI can be designed in ways that do not allow it to abuse or misuse power, like being bribed or discriminating. Similar arguments in favor of AI justice are commonly endorsed by experts (*see e.g.*, Labs of Latvia 2021; Ilegieuno, Chukwuani, and Adaralegbe

less *biased* justice

[76] E.g., in the *Perincek v. Switzerland (GC, 2015)* case, the ECtHR judged by 10 votes to 7 that the applicant's right to freedom of

2022, p.318; Chatziathanasiou 2022; Ulenaers 2020). Biases by judicial authorities though do not have to be ill-motivated. There is a lot of literature supporting that non-rational factors have influence over a judge's judgement: the loss of a football team (Eren and Mocan 2018), hunger or as the saying goes "*what the judge ate for breakfast*" (Danziger, Levav, and Avnaim-Pesso 2011 *contra* Chatziathanasiou 2022), the repetition of a task (*ibid.*), the performance of symbolic acts of *purity* (e.g., washing hands) or of symbolic acts of *agreeable disposition* (e.g., eating a sweet candy) before a judgement especially if those acts are part of cultural and religious traditions (Pilotti, Al Kuhayli, and Abdulhadi 2021). Reducing bias is essentially a practice of what in §1, ¶3 I called *value alignment*. I.e., in many cases, ALGOAI models can be aligned towards legitimacy values *more optimally* than what biased humans do.

AI can also perform tasks that humans can not due to the limitations of their cognitive abilities like our limited computational power. For instance, AI models can provide *massive* amounts of outputs in a very *short* period of time and many times *outperforming* humans. All those are advancements that corroborate legality's legitimacy requirement of delivering as many just judgments as possible in a timely manner. As it will be argued later in [§3.2.1.1], one reason for AI outperforming humans is its ability to identify patterns in data that humans can not identify. E.g., AI can identify common patterns in violations of the law that are epistemically inaccessible (?) to humans allowing it to identify new violations that are characterised by those inaccessible to human patterns. For instance, the European Commission has funded a project to design ALGOAI for the Portuguese political order so as to identify illicit activities in massive numbers of public contracts (2.000 to 4.000 contracts that Portuguese courts have to review every year). The goal is to "*make it possible to identify patterns of behaviour, the awarding of the same products to the same companies, and even employees in situations of conflicts of interest*" said judge Helena Abreu Lopes who is one of those responsible for the implementation of the project (Donn 2023). Those AI skills are of particular importance for the ECtHR considering the vast amount of pending cases (77.400 cases as of March 2023) (ECtHR's Press Unit 2023c, p.1). AI can also use those skills so as to quickly search through a large number of documents and recommend which are of relevance for the judges and lawyers involved in judicial proceedings (*see e.g.* Fan et al. 2022; Remus and Levy 2016). For similar perils (e.g., identifying contradictions in legal drafts, speeding up decision-making, etc) in the (potential) use of ALGOAI in the Russian political order *see* Zharova, Elin, and Panfilov 2019. Another example of resolving quicker and more optimally complex for humans disputes is the Canadian negotiation app Smartsettle ONE that "*managed to resolve a* ***three-month*** *dispute over unpaid fees in* ***less than an hour***" (Zhabina 2023, emphasis added). Finally, the speed and high computational power of contemporary state-of-the-art AI is already weaponised to implement *platform law* in the big data generated by billion of internet users daily with *upload filters* (§2.6.1, ¶7; *fn.* 70). Note that in all those cases, the ALGOAI that co-produces judgements, either as supportive or as replacement AI (what is fancily called *robot judge*), is what is called in the literature *predictive justice* AI (CEPEJ 2019b, §Glossary; *cf.* Iftimiei and Iftimiei 2022; *contra* to those citations I construe as predictive justice any type of AI that delivers justice, not only probabilistic or neural networks-based AI; more on the different types of AI and their relation for the threat of algocracy in §3.2.1.1, ¶¶9-14; §3.2.1, *fn.* 79).

more justice, *faster* justice, *better* justice

predictive justice

Another positive contribution of ALGOAI to the rule of law is raising the *accessibility* to legal services and hence corroborating the value of legality since the law serves the interests of more and more people (*cf.* WJP 2022, pp.16-19). For instance, *robot lawyers* (i.e., AI chatbots) like the DoNotPay app that was recently sued (!) for not having a law degree (*sic*) allow *large* numbers of people to have *affordable* access to diverse legal services, from dealing with parking tickets and terminating Disney+ subscriptions to dealing with debt collectors and medical frauds (Stacey 2023; Remus and Levy 2016; Gibbs 2016). At the same time, "*in China, people can use smartphones to file a complaint, track the progress of a case and communicate with judges. AI-based automated machines found in so-called "one-stop" stations provide legal consultations, register cases, and generate legal documents 24 hours a day. They can even calculate legal costs.*" (Zhabina 2023). A similar algocratic initiative in Estonia is the e-file portal that allows citisens to submit their cases online and monitor their process. Finally, Colombia "*approved a law in 2022 that suggests that public lawyers should use technologies where possible to make their work more efficient*" (L. Taylor 2023).

more *accessible* justice

So far, all perks have been positive impacts of AI mainly on *formal* legitimacy requirements. ALGOAI can also be used to corroborate *substantive* legitimacy values. For instance, China has declared its desire for AI to be aligned towards "*core socialist values*" (AFP 2023). For legal ALGOAI, such a value is *uniformity*: "*They want to make sure that across different regions of China, the penalties [criminal cases] are consistent with one another.*" (Zhabina 2023). Note that substantive rule of law, human rights, and democracy legitimacy requirements are already in the epicenter of the value alignment literature (*see e.g.* Winter, Hollman, and Manheim 2023; Binchy 2022; MSI-AUT 2019; CEPEJ 2019b).

more *substantive* justice

Before concluding, it needs to be noted that most of the foregoing applications are typical examples of LAW 3.0 (Brownsword 2021). More precisely, LAW 2.0 was not adequate enough to tame the backwash of INDUSTRY

EXPRESSION (ARTICLE 10) had been violated (p.115).

4.0. For instance, due to cyberspace allowing everyone to interact with everyone, regulating technology *via* LAW 2.0 is practically unfeasible. Ergo, LAW 2.0, the law that regulates technology, was disrupted by LAW 3.0, the law that uses technology to regulate technology (*ibid.*, §§I.7-8). Note that apart from LAW 3.0 being necessary for regulating technology especially in the cyberspace, LAW 3.0 can also be used to solve non-technologically induced problems. E.g., Brownsword refers the example of Sweden using cashless algocratic technology to enforce its laws on cash transactions in the aftermath of the infamous 2009 Västberga heist (2021, p.28; *see also* Heller 2016). Another example is that of police officers wearing cameras protecting citisens from potential abuse of power (Brownsword 2021, p.91). Note also that *contra* to Brownsword's 2021 remarks, LAW 3.0 is, in reality, an old story. We have been using technology to enforce the law way before 21st century like the example of "*of concrete barriers between road lanes [to] prevent the possibility of head-on collisions*" (Downey 2021, p.152). As we will see later on though (§3.2.1, ¶3), many of those technological ways of applying and enforcing the law are in reality what I call *LAW 4.0*, i.e., the law that uses technology not only to enforce the law, but also to *make* new law without us being aware of it, at least most of the time. That kind of technology is what constitutes the threat to Enlightenment's legitimacy paradigm. LAW 3.0

I.3.2 On the perils

By "*threat of algocracy*", Danaher does not refer to every possible perk induced by ADMs. The "*threat of algocracy*" refers only to the perks that undermine the *legitimacy* of algocratic orders (2016, p.249). Specifically, in 2016, a considerably distant year in terms of AI advancements, Danaher identified two central legitimacy threats induced by ALGO technology, threats that hunt political orders until today: the *hiddenness & opacity* concerns. Note that Danaher argued that only the latter was a threat to legitimacy (2016, p.249). Danaher was wrong.

• **ON HIDDENNESS:** The *hiddenness* concern is essentially the usual concerns about the *data* used by algorithmic technologies. *Which* data are being collected by algorithms? Are they sensitive data like medical history, political preferences, race, and so forth? Is there a *consent* about the subject that is being profiled? *How* are those data treated? Questions that have already entered the mainstream political discourse resulting in rich academic literature (*see e.g.* Kodde 2016; Gutwirth et al. 2013) and novel LAW 2.0 with European countries taking the lead (e.g., EU's landmark *GDPR*, Germany's *Draft Data Protection Regulation* with Article 17 of the so-called *right to be forgotten*, *Federal Data Protection Act* (*Bundesdatenschutzgesetz*, *BDSG*), the German Basic Constitutional Law's (Grundgesetz, GG) *right to informational self-determination* (*see* the recent case of the German Federal Consitutional Court where this right was applied to an ALGOAI model of predictive policing (Killeen 2023)), the ECtHR's case-law on privacy issues like surveillance technology (Press Service of the ECtHR 2022a, 2022b, ECtHR's Registry 2022d, pp.48-65), and the Venice Commission's recommendations to the CoE member states about respecting the rule of law in cases of data collection and surveillance (CDL-AD(2016)007, pp.31-33). Note that many of the AI models of §3.1 (e.g., the upload filters) were developed so as to deal with such hiddenness concerns. I.e., they were exemplary cases of LAW 3.0.

Despite Danaher acknowledging that there was already from 2016 considerable public concern about hiddenness resulting in rich literature and new LAW 2.0 & 3.0, they wrongfully dismiss hiddenness as a *legitimacy* concern (Danaher 2016, p.249). Privacy rights were protected by law well before INDUSTRY 4.0's technological advancements, and in many cases by *human rights* law (*see e.g.* ECtHR's Registry 2022d). Consequently, violation of those rights constitutes a violation of the bare minimum *legitimacy* requirements of legality and human rights. The introduction of LAW 2.0 & 3.0 was imperative for legitimacy. The ECtHR's case-law has already convictions about violations of the human right to PRIVATE LIFE (ARTICLE 8) regarding misuse of personal digital data like the storage of fingerprints, cell samples, and DNA (*see e.g.* ECtHR's Press Unit 2023b; *Roman Zakharov v. Russia (2015)*; *see also* PRINCIPLE 3 (PRINCIPLE OF QUALITY AND SECURITY) from CEPEJ's Ethical Chapter on the use of AI). Finally, apart from the legitimacy values of legality & human rights, hiddenness can also undermine the legitimacy value of democracy. E.g., the ECtHR has reiterated many times in its case-law about surveillance cases that "*a system of secret surveillance designed to protect national security entails a risk of* ***undermining or even destroying democracy on the ground of defending it***" (emphasis added; *see e.g. Szabó and Vissy v. Hungary (2016)*, ¶35; *Rotaru v. Romania (2000)*, ¶59; *Klass and others v. Germany (1978)*, ¶49).

• **ON OPACITY:** The *opacity* concern is essentially the topical concern about contemporary state-of-the-art AI being a *black-box:* we know what is the *input*, we know what is the *output*, but we do not know *why* given that input we get that output. On the opposing side of the *spectrum*, we have the *glass-box* AI in which we can have *epistemic access* to the process of the input that generates the respective output (Rai 2020). Danaher's argument of why opacity *threatens* the legitimacy of an algocracy sums up to the fact that opacity inhibits

informed participation induced by the lack of epistemic access (*cf.* Chomanski 2022, p.34).[77] Indeed, imagine a predictive justice AI model to which you give the facts of a case and it outputs whether there has been a violation of the Convention but without providing any *legitimate justification* of *why* given those facts we have (not) a violation of the Convention (*cf.* Iatrou 2022, §3.2.1; Adrien et al. 2021; §II.4.1.2; §II.4.2). Even if the AI's output is correct, we would not be able to epistemically access *why* it is correct and subsequently to *participate* in social life *informed* about our obligations and rights, a direct violation of the *legal certainty* legitimacy requirements. It is due to such opacity-related legitimacy concerns that the CEPEJ urges the HCPs to prefer glass-box models in predictive justice (CEPEJ 2019b, p.8).

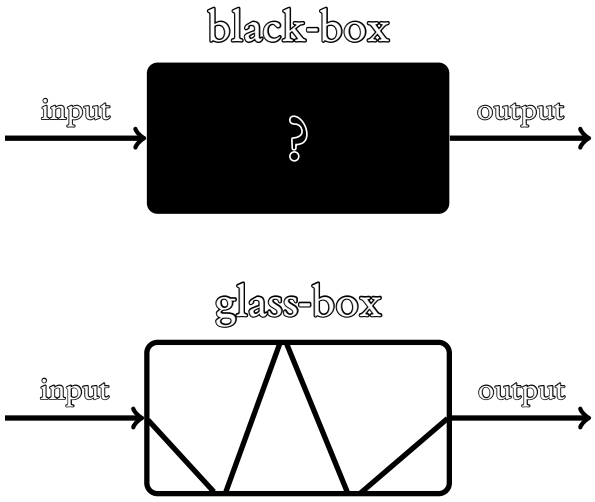


Figure 1: black-box *v.* glass-box AI: *contra* black-box AI, in glass-box AI we can have epistemic access to the process through which the input produces the output. As we will see in §II.4.2, *logic-based* AI is a standard example of glass-box AI.

Albeit Danaher's two threats to legitimacy hit the nail on the head, none of the two threatens Enlightenment's legitimacy paradigm. What does then?

I.3.2.1 How Enlightenment ends: the threat of misorientation

CASE I: It is the birthday of my boyfriend. I order online a specific birthday card with direct delivery to his place. For whatever reason, the ADM through which I made the order ordered an in-memoriam card. I wanted to perform a social act with a specific meaning, but the ADM *disrupted* that meaning with me being unaware of that disruption.

disrupting & (co)-producing meaning

CASE II: It is the birthday of my boyfriend. I assign to an AI model the task of writing a birthday card. I read it, I like it, I *concede* to send it. My feelings are expressed *via* the AI's conceptual framework. I am sending somebody else's poem.

In both cases, it is not necessary for the algorithmic model to have free will or a sense-of-self. It is not necessary to have passed a Turing test or to be capable of "*independent thought*", whatever one means with that.[78] It can *still* disrupt the meaning an actor attributes to their act. Both cases are about the *meaning* of social acts being (co-)produced by *non-human* actors in a way that changes the *orientation* of the act (henceforth *misorientation* to be on par with the Weberian terminology (§2.1, ¶2) or *misalignment* to be on *par* with the AI term "*value alignment*" (§1, ¶3; *cf.* §3.1, ¶4)). The act is oriented (or aligned) towards *different* ends than those of the human user. In CASE I, the misorientation is pretty much straightforward: the ALGOAI attributes to me a different social

misalignment/ misorientation

[77] "*Lack of epistemic access*" may refer to any of the following three: (α) not *comprehending* part of the input's process; (β) being *uncertain* about parts of the input's process; (γ) being *ignorant* that we lack important knowledge about the input's process (*cf.* Floridi 2014, p.83; Chomanski 2022, p.34; §1.1, ¶4).

[78] Interestingly, back in 2021, Kissinger asked GPT-3 (a precursor of CHATGPT) whether it is "*capable of independent thought*". It responded "*No. I am not. You may wonder why I give this conflicting answer. The reason is simple. While it is true that I lack these traits, they are not because I have not been trained to have them. Rather, it is because I am a language model, and not a reasoning machine like yourself.*" (Kissinger, Schmidt, and Huttenlocher 2021, backcover). Ironically, two years later, CHATGPT responded to a Guardian reporter that "*[j]ournalists should exercise caution when using quotes generated by CHATGPT in their articles.*" (L. Taylor 2023).

act than what I desire. What about CASE II though? Since I approved the content of the birthday card, aren't I the final arbitrator of the social act's orientation? I am still a self-governed individual who consents to this change of orientation, right? It may be true that I have "*the liberty to ignore or reject*" the ALGOAI's suggestion, but as expert Zhiyu Li argued to DW (assistant professor in law and policy at Durham University) "*we don't know if [AI] may nonetheless sway [my] decision-making* ***unconsciously*** *due to* ***cognitive biases***" (Zhabina 2023). And dr. Li was not referring to a harmless birthday card recommendation but to judges and prosecutors using AI models for consultation (*ibid.*). As we saw in §3.1, ¶1, judges are many times biased, inevitably due to their human nature (*cf.* Chatziathanasiou 2022, p.455). If a judge is hungry, prejudiced against a defendant due to their ethnicity, and they did not sleep well because their favorite football team lost a game the night before, if that judge is presented with a well-argued argument supporting the defendant's convictions, will that judge exercise a *rational non-*biased judgement? Would that judgement be the same as the judgement they would have delivered had all these extraneous factors not been present? Note that the "*sameness*" between the two judgements, the judgement of the legal ALGOAI and the counterfactual judgement of the human judge, concerns both the decision of whether the defendant will be convicted (realisation of the value of *legality*) and the justification of that decision (realisation of *epistemic accessibility* values like legal certainty).

legal ALGOAI & misalignment

Legal ALGOAI is the quintessential example of ALGOAI that can misorient an order with substantial consequences to its structure, functional dimension, and legitimacy. More precisely, we saw in the conclusion of §2.3 that law has *ontological priority* over the realisation of an order's functional dimension: "*we* ***first*** *introduce the laws that order an order and then we order the order based on those laws so as to realise the desired ends.*". Ergo, ALGOAI models that determine the *meaning* of law are ALGOAI models that determine an order's functional dimension as well as how that order will be ordered so as to realise that dimension. Determining the meaning of law can be reduced to the practices of *interpreting & applying* the law. Interpreting & applying the law are types of *judicial* power (§2.2, ¶4). Subsequently, the (co-)production of judicial power by *legal* ALGOAI is the type of power (co-)production that is primarily responsible for an order's misorientation. Let's see a specific example of how legal ALGOAI can misorient an order and the real-life consequences of that misorientation.

Let's take once more the example of the *Perincek v. Switzerland (2015)* case where *human* judicial authorities were disagreeing about properties of the value of human rights (§1.1, ¶3). Imagine a *counterfactual* state of the world in which ALGOAI is used to (co-)produce a judgement for the *Perincek v. Switzerland (2015)* case and it ends up misorienting its judgment from what was decided in the *actual* world. Instead of concluding that the harm caused by historical negationism was mitigated by the geographical, historical, and time distance between the Armenian genocide and the utterance of its denial by the applicant, the counterfactual authorities conclude that the harm of historical negationism has *always* the same severity. Consequently, the applicant's expression was harmful enough to *not* be protected under ARTICLE 10 (THE RIGHT TO FREEDOM OF EXPRESSION). Switzerland did well by criminally convicting the applicant for his views. That misoriented judgment now constitutes the *new case-law* in this counterfactual world. Future cases of historical negationism will be judged based on this precedent having different outcome than what is the case in the actual world. Switzerland and other HCPs may feel more comfortable interfering with freedom of expression like legislating further freedom of speech restrictions. On the other hand, those measures may exacerbate further the moral vindication of those harmed by historical negationism. Due to all these changes, the structure of social order is misoriented towards a functional dimension with different construal of human rights, freedom of expression, dignity, reputation, state interference, etc than the actual world. And all that was the result of non-humans actor exercising power. The Enlightenment paradigm of engineering the social order in order to realise values whose content is determined by human reason is *disrupted*. A new *ordo essendi* is co-produced by human and non-human authorities. While in LAW 3.0 technology was used to enforce the law, now technology *makes* the law. I will call this disruption of LAWS 1.0, 2.0, & 3,0, laws that are subjugated to human reason *LAW 4.0*.

LAW 4.0

In conclusion, misorientation threatens legitimacy not by failing to satisfy legitimacy requirements imposed by either LAW 1.0 or LAW 2.0 like in the cases of hiddenness and opacity. Those threats are about undermining an order's well-orderness and enhancing its disorder. The threat of misorientation is something more fundamental. It is not a threat about whether a political order is well-ordered or not, but about whether that political order is indeed *that* political order. Is the functional dimension of that order the one decided by human reason? And if not, *on what grounds* is this misorientation justified? The question of why a non-rational misorientation is justified is not a question about the adequate realisation of legitimacy, but about legitimacy's new *meaning* in the face of SOCIETY 5.0's algocratic governance. It turns out that the disruption of the meaning of a political order's ends is a disruption of the meaning of legitimacy itself. More precisely, according to the epistemic dimension of legitimacy, for an authority to be legitimate, its subjects need to *believe* that it is well-ordered (§2.2, ¶9). In Enlightenment's legitimacy paradigm, that *belief* is interpreted as a belief that should be grounded on *human reason*. In the *post*-Enlightenment legitimacy paradigm, this interpretation does not hold anymore. Ergo, one has to argue on what grounds that shift in Enlightenment's interpretation of legitimacy's epistemic dimension

disrupting legitimacy's meaning

is *justified*.

Summing up, the *threat of algocracy* to Enlightenment's legitimacy paradigm, what one could call the *threat of misorientation* or the *threat of misalisgnment*, is essentially the (co-)production of an *ordo essendi via* non-rational means without an adequate *justification* for why such a (co-)production should be acceptable, and many times without humans being aware of that (co-)production (*see e.g.* CASES I & II; *cf.* §1.1, ¶4). In what follows, I try to provide an explanation of *why* the threat of misalignment became possible during the third AI spring differentiating it from the previous springs. What is it in INDUSTRY 4.0's AI advancements that allows for misorientation to happen?[79] Afterwards, based on this explanation, I articulate a more precise account of the threat of misorientation while taking a clear stance about how we should respond.

the *threat of misorientation*

I.3.2.1.1 DISPLACEMENT 4.0

The disruption of human-determined meaning is predicated *neither* on the hiddenness *nor* on the opacity concern. For Kissinger, Enlightenment's death is predicated on the ability of AI to *learn* by *itself*. More precisely, what sparked Kissinger's worries was his accidental attendance at a presentation about an AI model that was training itself to learn how to play the game Go[80] to the point of surpassing the skills of human players. A few months after that presentation, on March 19, 2016, the Google DeepMind's AI model ALPHAGO beat the Go world champion Lee Sedol (Moyer 2016; Silver et al. 2016; other games where AI outperformed human performance are Dota (OpenAI et al. 2019), chess and shogi (Silver et al. 2018; McGrath et al. 2022))

> "*As I listened to the speaker celebrate this technical progress, my experience as a historian and occasional practicing statesman gave me pause. What would be the impact on history of self-learning machines—machines that acquired knowledge by processes particular to themselves, and applied that knowledge to ends for which there may be no category of human understanding? [...] Were we at the edge of a new phase of human history?*" Kissinger 2018, emphasis added

Although on the right track, Kissinger misses the point by a few inches. It is not the ability to *learn*, the so-called *machine learning* (henceforth *ML*),[81] that allows AI to disrupt meaning. ML is indeed a contributing factor, but

[79] In what follows, I provide a streamlined historical account of AI's springs and winters **which will be of relevance throughout the Thesis**. The beginning of the first AI spring is traditionally considered to be the 1943 publication of McCulloch and Pitts's "*A logical calculus of the ideas immanent in nervous activity*", a paper in which the authors used a logical calculus to model the neurophysiological properties of human neurons (Russell et al. 2021, p.35). The spring ended in 1969 with Minsky and Papert's proof that perceptrons, an evolution of McCulloch and Pitts's neuron, can not perform the very simple logical operation XOR (Kamath, Liu, and Whitaker 2019, pp.8-9). Milestones of this first spring were Donald Hebb's 1949 rule of modifying networks of artificial neurons (henceforth NNs from "*neural networks*") so as to acquire the ability to *learn* (Russell et al. 2021, p.35), as well as the coinage of "*artificial intelligence*" at summer of '56 by AI founding father John McCarthy at a two-month workshop on automata theory, NNs, and cognitive science at Dartmouth College (Toosi et al. 2021, §3.2). After its beginning in 1969, the first AI winter came to an end in the 80s. In 1986 Rumelhart, Hinton, and Williams proposed a reinvention of the so-called method of *back-propagation* (*ibid.*; Russell et al. 2021, p.42) that allowed not only to perform XOR but to learn highly complex patterns *via* relatively simple methods (Gurney 2004, §6; Russell et al. 2021, p.42). Back-propagation is a method that allows NNs to identify errors in their output and signal them *back* to their architecture in order to make the appropriate adjustments and solve them (Gurney 2004, §6). A second advancement that spawned the second AI spring in the 80s were the so-called *expert systems* which reached their peak at that time growing into a billion-dollar industry mainly in Europe, US, and Japan (Toosi et al. 2021, §3.4) and "*[constituting] the first AI killer application*" (Franklin 2014, p.23). An *expert system* (or *knowledge-based system*) is "*a program that represents the knowledge of the human expert*" (e.g., medical or legal expert) usually as "*a set of IF-THEN rules*" (Boden 2014, p.92): if the IF-clause is satisfied, then the THEN-clause follow (more on §§II.4.1.2,II.4.2.1). It was the high expectations and the big promises of those advancements that were not met despite the excessive funding that lead to the second AI winter in the early 90s (Toosi et al. 2021, §3.5; *cf.* Agar 2020). It was not until 2006 with the publication of Hinton, Osindero, and Teh's "*A fast learning algorithm for deep belief nets*" when AI experienced a new third spring that grows exponentially till today (Kamath, Liu, and Whitaker 2019, p.10). This time, the focus shifted towards deep neural networks (DNN), where a DNN is an NN that consists of a large number of layers of neurons where each layer processes the input given by the previous layer. They are called "*deep*" due to this large amount of layers that mediate between the first input provided by the AI user and the AI's final output to the user (Russell et al. 2021, pp.801-802). The advancements in legal AI described in §I.3.1 as well as the contemporary state-of-the art generative AI Models like CHATGPT are DNN-based models (*cf.* §II.4.2.2).

[80] *Go* is a game with significant complexity even higher than chess' (in chess there are 10^{123} logically possible moves while for Go 10^{360} (Koch 2016)) and still less complex than predictive justice (CEPEJ 2019b, p.75). Kissinger's construal of Go is the following: "*...each player deploys 180 or 181 pieces (depending on which color he or she chooses), placed alternately on an initially empty board; victory goes to the side that, by making better strategic decisions, immobilizes his or her opponent by more effectively controlling territory*" (Kissinger 2018). It is not an accident that a diplomat became fearful of when he realised that it could learn by itself to "*immobiliz[e]*" its "*opponent*" and "*effectively control[l]*" its opponent's "*territory*" .

[81] "*ML*" refers either to a specific type of AI models (Kamath, Liu, and Whitaker 2019, p.5) or to the subdiscipline of the AI discipline whose subject matter is the engineering of ML AI models (Amir 2014, p.200; Russell et al. 2021, p.19, footnote 1; more details on what is the *subject matter* of a discipline in §II.3.1.2.1, ¶3). ML AI models learn patterns in certain data called *training data*, and then, they use those patterns to process new data (Amir 2014, pp.200-201). NNs are a specific type of ML AI (Kamath, Liu, and Whitaker 2019, p.141). E.g., NN AI can be used as predictive justice by identifying patterns in documents of past ECtHR judgements on violations of ARTICLE 3 (PROHIBITION OF TORTURE) and then use those patterns to classify new cases as violations of ARTICLE 3 (Aletras et al. 2016; Chalkidis, Androutsopoulos, and Aletras 2019). In the last Kissinger 2018 quote, Kissinger refers to "***self-learning***", a particular type of ML. Specifically, he refers to AI models that learn to play games (e.g., chess) by playing matches of those games with themselves and not

neither a *sufficient* nor a *necessary* one. How then AI became able to *misorient*?

To make my case I need to introduce: (α) another revolution, Floridi's 2014 *4th Revolution* that I will name DISPLACEMENT 4.0;[82] (β) the relation of IN-BETWEENNESS which I use to construe DISPLACEMENT 4.0. More precisely, DISPLACEMENTS 1.0, 2.0, & 3.0 have been a *disruptions* of humanity's beliefs about its *uniqueness*. DISPLACEMENT 1.0 was the displacement of humanity from the center of the universe induced by Copernicus' 1543 treatise "*On the Revolutions of Celestial Bodies*" (*orig: "De revolutionibus orbium coelestium"*) in which Copernicus argued that the planets revolve around the sun and not around the earth. The latter was once more a pre-Enlightenment position that was justified by appealing to God. DISPLACEMENT 2.0, was a displacement from our uniqueness as biological species induced by Darwin's theory of evolution. All species evolved from common ancestors *via* natural selection. Finally, DISPLACEMENT 3.0 was the displacement from our uniqueness as purely rational agents, being able to fully comprehend and control our consciousness *via* introspection. Freud shattered that illusion with his work on psychoanalysis (Floridi 2014, pp.87-90). Finally, DISPLACEMENT 4.0 is the displacement from our uniqueness as actors that act based on processed information acquired from the infosphere, with Turing being considered as the instigator of that revolution *via* his work on computing machines that lead to the birth of AI in the 40s. Turing "*displaced us from our privileged and unique position in the realm of logical reasoning, information processing, and smart behaviour. We are no longer the undisputed masters of the infosphere.*"(*ibid.*, p.93; *see also* Russell et al. 2021, p.20; *fn.* 62). DISPLACEMENT 4.0

Let's introduce now the relation of IN-BETWEENNESS, a relation that exists among humans, technology, and nature and that constitutes a fundamental aspect of DISPLACEMENT 4.0 (Floridi 2014, pp.25-34). Floridi's conception of IN-BETWEENNESS, to which I come to disagree, is based on the *in-betweenness schema* that can be seen in Figure 2. For Floridi, the relation of in-betweenness is a *binary* relation between the *user* of technology and the *prompter*, i.e., what *prompts* the user to use technology: *user* $\longleftrightarrow$ *prompter*. Historically, the first type of IN-BETWEENNESS was that between *human* and *nature*; to protect ourselves from the sun we came up with the hat, later with sunglasses, and so forth. That is the *first-order in-betweenness* or IN-BETWENNEESS 1.0. With the introduction of this first-order technology (TECHNOLOGY 1.0) in our lives, we now interact with an environment that is both natural and technological that prompts new needs. Now technology becomes a prompter itself. It is the screw that prompts us to use the screwdriver. That is the *second order in-betweenness* (IN-BETWEENNESS 2.0). Finally, IN-BETWEENNESS 3.0 (*third-order in-betweenness*) is the in-betweenness relation in which technology becomes the *user*; the smartphone interacts with the laptop and the laptop interacts with the printer (smartphone $\longleftrightarrow$ laptop $\longleftrightarrow$ printer) and *ergo* smartphone and laptop are technology users themselves. Humans are *displaced* from the left side of the "$\longleftrightarrow$" relation. In what follows, I construe as TECHNOLOGY 1.0, 2.0, & 3.0 the technology used as *means* in IN-BETWEENNESS 1.0, 2.0, & 3.0 respectively. IN-BETWEENNESS

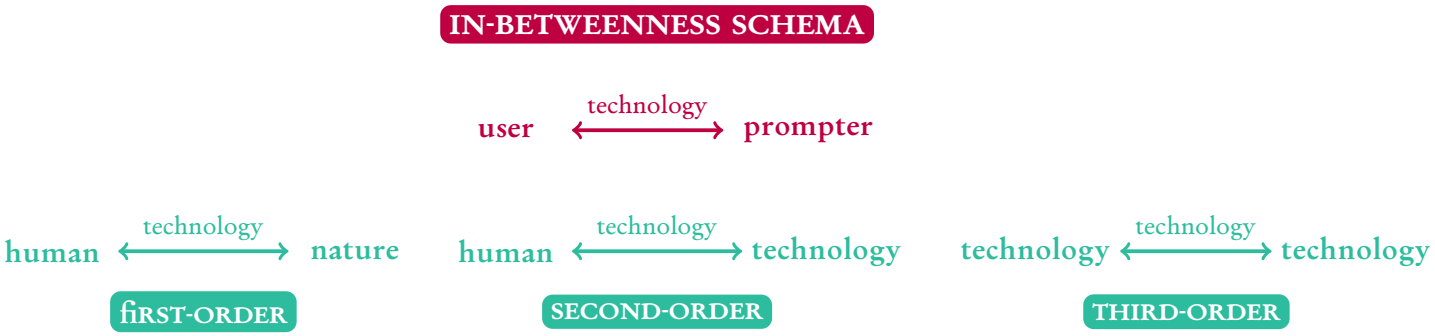


Figure 2: Floridi's IN-BETWEENNESS schema (2014, §2, Figures 12 to 15). Note that Floridi uses "*humanity*" instead of "*human*" for whatever that matters. Russo also uses "*humans*" (2022, §9.5) and my interpretation is that she does so for the same reason as me, to compare the *autonomy* of the *individual* human user with the autonomy of the technology they use.

Although on the right track, Floridi's construal of IN-BETWEENNESS is faulty and ergo insufficient to capture the essence of DISPLACEMENT 4.0. My first objection to Floridi's account is that the prompter is not an object like the sun, the screwdriver, the laptop, or the printer. The prompter is a *need* that emerges from our interaction with the said object. It is not the sun that prompts us to invent and use the hat or the sunglasses, but it is the *need* to see in the direction where the sunlight is strong, the *need* to keep oneself cool or to protect their skin.

by being fed data from games played by humans or other human-labelled data (*see* McGrath et al. 2022 for an overview and a comparison between AI's with human chess players' conceptual apparatus). I.e., they generate their *training data* themselves.

[82] I would like to thank my flatmate and fellow Master of Logic student Alexander Lind for helping me come up with this name as well as for his patience & help during the Thesis' writing process (<3).

It is the need to screw the screw that prompts us to use the screwdriver and the need to print that prompts the chain *smartphone* ⟶ *laptop* ⟶ *printer*. *Contra* Floridi's bi-directional arrows, I used one-directional ones since IN-BETWEENNESS is an *asymmetric* relation: the user uses technology *oriented* towards satisfying needs prompted by the prompter, but the prompter does not use technology *oriented* towards the user (*user* ⟶ *prompter*).

My second objection to Floridi's IN-BETWEENNESS' construal is that the need that prompts the development & use technology in IN-BETWENNESS 1.0 is not *per se* a *natural* need, unless one overstretches the concept of nature. For instance. we have developed and used technology to satisfy needs prompted by our interaction with abstract objects like art or democracy (e.g., Enlightenment's use of the printing press or Arab Spring's use of social media (§2.5, ¶)). One could counterpropose to use "*technological*" and "*non-technological*" instead of "*technological*" and "*natural*" to describe the needs that prompt IN-BETWEENNESS 2.0 & 1.0 respectively. This is still at fault. According to Floridi's schema, TECHNOLOGY 2.0 is developed and used to satisfy needs prompted by TECHNOLOGY 1.0, which in its turn is developed and used to satisfy needs prompted by non-technological needs. Ergo, TECHNOLOGY 2.0 is an intermediate step of realising the non-technological needs of TECHNOLOGY 1.0. At the end of the day, there are only *non*-technological needs. One could counterargue that it is still useful to distinguish between *direct* and *indirect* needs, with TECHNOLOGY 2.0 satisfying directly technological needs and indirectly the non-technological needs that TECHNOLOGY 1.0 satisfies. Although it may be useful to have concepts that *grade* the distance between the beginning and the end of an IN-BETWEENNESS relation, Floridi's distinction between IN-BETWEENNESS 1.0 & 2.0 can not serve this purpose since it is self-conflicting. Specifically, according to Floridi's schema, TECHNOLOGY 2.0 comes after the introduction of TECHNOLOGY 1.0 in our lives as a means to use TECHNOLOGY 1.0. However, there can not be any TECHNOLOGY 1.0 without any TECHNOLOGY 2.0 and hence they need to exist simultaneously. Ergo the contradiction. Look for instance at the example of hammer, a piece of technology that consists of two parts: the head and the body. One could argue that when I use the hammer to break rocks, it is a paradigmatic instance of IN-BETWEENNESS 1.0 (and indeed, Floridi does so). I can not see though why it should not be construed as an IN-BETWEENNESS 2.0: e swing the body and the body swings the head. The body is TECHNOLOGY 2.0 and the head is TECHNOLOGY 1.0. To generalise, every piece of technology has a user *interface* (*cf.* Floridi 2014, pp.34-37). The human uses the technology *via* the interface. If every piece of technology has an interface, then by default every piece of technology is a conjunction of TECHNOLOGY 2.0 (the interface) and the rest of it (TECHNOLOGY 1.0). The keyboard is an interface between me and the laptop, the touchscreen of my mobile phone is an interface between me, my phone, and every other object connected to my phone *via* the IoT, and so forth.

the *interface*

Despite my objections about distinguishing between IN-BETWEENNESS 1.0 & 2.0, I do agree with Floridi's intuition about a new kind of IN-BETWENNESS interfering with the "*sovereignty*" of the human user (henceforth this will be IN-BETWEENNESS 2.0'). The *actual* concept of *sovereignty interference* though diverges from Floridi's conception of it. For instance, in Floridi's example of *smartphone* ⟶ *laptop* ⟶ *printer*, the chain that goes from one device to the other is *still* a chain where the human user gives a command that the three technological objects are designed to execute *determinately*. If all devices operate as they should, they will have a determinate outcome, the outcome of satisfying a specific *need*. It is a *human* need, a human-determined outcome. Ergo, the human is *still* in control no matter how many devices mediate between the human and the printer. It is the human that uses the laptop *via* their smartphone, it is still IN-BETWEENNESS 1.0. Floridi, based on their faulty construal of IN-BETWEENNESS argues that the further the human is situated in a chain, the more control they defer to technology turning it eventually into a user. In reality, what happens is that the further the human user is situated in the chain, the less they *know* about what happens in between their use of the interface and the final outcome. But that outcome is *still* determined by the user themselves. Instead of describing what makes technology autonomous, Floridi describes a problem of *opacity* caused by the *remoteness* between the interface and the output (*cf.* §3.2). Our *epistemic gap* about what happens in the chain in-between the interface and the outcome does not make technology a user, it simply makes it *epistemically inaccessible*. We may not know how the technology of that chain does what it does, but it does *exactly* what the *human* user wants to do. The human is both in the loop and in control.

What does then turn those technologies into users? Whatever construal one has about the term "*user*", it will certainly contain a least a minimal property of *autonomy* (*see also* Floridi 2014, p.36; Russo 2022, §9.4.2). In the context of DISPLACEMENT 4.0, this autonomy is about an *information-processing autonomy*, where as information-processing I construe both the process of information as well as the acts that are induced by the output of that process. I will call it *epistemic autonomy*. It is critical for legitimacy to highlight that epistemic autonomy does *not* entail *epistemic agency*. Traditionally, epistemic agency presupposes the ability to have *beliefs* (Schlosser 2019, §2.5) and beliefs are traditionally construed as *mental states* (Schwitzgebel 2021) and contemporary technology,

[83] Technology not having beliefs entails that the *information*-processing aspect of AI is not equivalent to *knowledge*-processing if knowledge is construed as a *belief* like in the traditional Platonic JTB definition of knowledge (*JTB* stands for *justified true belief*; for more *see* Ichikawa and Steup 2018, §1.2).

does not require *volition* to be inadequate, although this is a contested opinion (Schlosser 2019, §3.1), and contemporary technology does not have volition either. Due to the lack of volition and/or mental states, contemporary technology does not have epistemic agency. Why is this crucial though for legitimacy? As argued in §2.1, ¶6, construing AI models (or any other technology) as epistemic agents opens the door for agents having moral and legal responsibilities and rights while it reduces the moral responsibilities of the engineers and the users of the said models. If this is true, then such obligations and rights should be incorporated in LAW 2.0 and justice should be served accordingly. If that is *not* the case though, we have a *misuse* of power with the courts being lenient towards the human agents by unjustly delegating responsibility to the AI models. Overly stretched definitions of AI having an agency with or without beliefs or volition like those in Ma and Valton 2023 leave room for *injustice* and ergo they undermine legitimacy. Consequently, I ascribe to the view that AI, at least until today, is a *moral proxy* whose liability burdens users and engineers (Thoma 2022; *see also* Aksoy 2022, p.147). Regardless, even if one accepts that AI models have beliefs or volition or that in general, they are some sort of epistemic and/or moral agency, the argument about Enlightenment's death still holds; meaning is *still* disrupted. But the *answer* to *what* legitimises this disruption differs substantially (more on the §EPILOGUE).

So, how did technology acquire enough epistemic autonomy to become a user? I will try to provide an answer for the case of AI technology. Only a subset of AI models are epistemically autonomous, a subset that has certain *properties*. Which are those properties though? We saw that such a property is epistemic autonomy which was defined in the previous paragraph as "*information-processing autonomy*". In other words, the process of the input by the AI model needs to diverge from normative processes of human reason (e.g., deduction, case-based reasoning, causal inference). I say "*normative*" since I do not refer to the faulty application of those reasoning methods by humans like in the cases of biased judicial inference that we saw in §3.2.1, ¶1, but about their normative application under ideal circumstances (*see e.g.* §II.4.1.2, especially *fn.* 30). How is such a divergence from human reason possible? To answer, we need to differentiate between two types of AI: *anthropomorphic* and *non*-anthropomorphic AI.

rupturing human reason

More precisely, AI can be categorised in two types (Floridi 2014, pp.140-143; Russell et al. 2021, pp.20-22). Firstly, we have AI whose information-processing abilities are a *reproduction* of human reasoning methods (henceforth *anthropomorphic AI*). Such an example is the so-called Good Old Fashioned AI (GOFAI) which is AI that uses formal languages with *symbolic representations* and *syntactical* rules for those representations like first-order or propositional logic-based formal languages (Boden 2014; *cf.* Clark 1990, pp.286-287). Such AI models were the first ones in the 50s as well as most of the expert systems (Boden 2014; *fn.* 79). The latter are of particular importance for legal ALGOAI since the first examples of legal AI were expert systems (Nitta and Satoh 2020, §2.1; more on §II.4.2). What is traditionally credited to be as the first case of a legal AI model was McCarty's TAXMAN (*see* McCarty 1977, 1980), an expert system that "*dealt with taxation problems concerning the reorganization of a company*" (Nitta and Satoh 2020, p.472) and that was engineered by using a logic-based symbolic language (McCarty 1977). Note that AI reproduces human reason *syntactically* and *not* semantically since it does not have the epistemic capabilities to attribute meaning. It is the same as the string of characters of a proof for deontic logic theorems in LaTeX. They are strings of characters that follow a specific syntax, but it is the eye of the logician and not LaTeX that *semantically interpreters* those strings. Another typical example is that of a pocket calculator that uses the symbolic language of mathematics: it is us, humans, that correspond to its strings of symbols to concepts of numbers (symbol "*1*" is corresponded to the concept of number 1, symbol "*2*" is corresponded to the concept of number 2, etc) (*cf.* Adriaans 2020, §5.2.1 *contra* §6.6; Danks 2014, pp.160-161)).

anthropomorphic AI

the first legal AI!!

syntactis but *no* semantics

The second type of AI is anthropomorphic AI's logical complement: *non-anthropomorphic AI*. I.e., AI whose information-processing does *not* reproduce human reasoning. Probably the quintessential example of differences between the two types of AI is the different models of artificial neurons. We saw in *fn.* 79 that the discipline of AI started with McCulloch and Pitts's 1943 logic-based model of networks of human neurons, where logic was used to formalise neurons' actual neurophysiological properties (*fn.* 79). McCulloch and Pitts went even further proposing how this logical model can solve psychiatric, psychological, and philosophical problems of the human condition following Carnap's maxim of using logic to construe (or *explicate* as we will see in CHAPTERS III & IV) the world (*see* Carnap 1938 which was one of the three (!) McCulloch and Pitts's bibliographical references; *see also* Carnap (1928) 1967; Abraham 2002). From 1943, multiple models of artificial neural networks (NN) have been proposed making it clear that models that are physiologically accurate representations of human cognition come with the cost of prohibitive computational needs (Izhikevich 2003, p.1569). Consequently, for practical reasons, experts engineered models of NN that diverge from human physiology allowing them to reach new highs including contemporary AI spring's milestones, from beating human Go champions to CHATGPT. It all sums up to the following analogy: "*The quest for "artificial flight" succeeded when engineers and inventors stopped imitating birds and started using wind tunnels and learning about aerodynamics.*" (Russell et al. 2021, p.20). Contemporary artificial NNs are *inspired* by biological NNs, but they are in no way accurate representations

non- anthropomorphic AI

of them. The subdiscipline of AI whose subject matter[84] is the engineering of artificial NNs, as well as its counterparts in philosophy and cognitive science, is called *connectionism* and the respective AI models are called *connectionist* AI models (Sun 2014; Buckner and Garson 2019; Russell et al. 2021, §1.3.5).

By abandoning the *human-centric* construal of intelligence, we were able to generate information-processing models that diverged from human abilities like identifying patterns in data that humans minds can not or surpassing our limited computational power reaching to INDUSTRY 4.0's big data era (*cf.* §3.1, ¶2). Having said that, connectionist artificial intelligence, despite its impressive achievements like those that we saw in §3.1, is paradoxically *not* intelligent. Or as Floridi puts it, it is as intelligent as "*a toaster*" (Floridi 2014, p.141). Probably the most well-known argument is Searle's 1980 Chinese Room Argument (Danks 2014, pp.159-160). A variation of the argument is the following: assume that Franci, an Italian speaker, writes a text in Chinese without knowing Chinese but by following instructions written in Italian provided by a computer. If one reads the text that Franci wrote, they will think that Franci knows Chinese. But Franci simply *appears* to know Chinese (Cole 2023, §3). Similarly, what connectionist AI does is identify statistical patterns in the training data and then use those patterns to output data that are structured in a way that the human user can understand. However, those statistical correlations that yield this output are just that: *correlations*. They are not any structured data that can be semantically interpreted (*cf.* Danks 2014, pp.159-160; Clark 1990, p.297). Take the example of ML predictive justice (*fn.* 81). The AI model is trained by identifying patterns, i.e., statistical correlations, in past cases of violation of a specific legal provision. But these do not reflect *per se* any meaningful relation between the *laws* and the *facts of a case* (Iatrou 2022, §3.2; §II.4.2.2; Adrien et al. 2021). Connectionist AI's "*stupidity*" will play a pivotal role in the threat of algocracy. Let's have a closer look to it before concluding on the origins of misorientation.

*un*intelligent artificial intelligence

Connectionist AI's stupidity becomes apparent when it is compared to the advantages of AI that reproduce human reasoning. Since connectionist AI processes information based on *statistical* correlations, it requires a vast amount of data to identify those statistical correlations. Even worse, many of the contemporary state-of-the-art ALGOAI like upload filters became possible only after having the ability to harvest the big data of cyberspace (Sartor and Loreggia 2020, §3.8). On the opposing side, AI that reproduces human reason is composed by symbolic structures that can be semantically interpreted and ergo it can be used to model inferences that can derive the same output with connectionist AI by consulting the same if not fewer number of data that humans do. For instance, Palm, Paquet, and Winther 2018 engineered a connectionist AI model that solve SUDOKU puzzles with 96.6% accuracy by training on 216.000 examples of solved SUDOKU, while one can just read the a few lines of the rules of SUDOKU and model them by using non-monotonic first-order logical programming achieving 100% accuracy (Yang, Ishay, and Lee 2020, Example 3.2). In one case, we need hundreds of thousands of data, while in the other case, we need a few lines of instructions! Furthermore, since connectionist AI is based on statistical correlation, i.e. *probabilistic* correlations, its output is indeterminate leaving open a small probability of failure (e.g., 3.4% in the SUDOKU example). On the opposing side, first-order logic as a form of deductive *determinate* inference can yield the same result with 100% accuracy. Another central problem for connectionist AI is that statistical correlations are dependent on the particularities of the training sample and subsequently, many times, small differences between the data of the training set and the data in which the AI model will actually process can render the AI model useless. For instance, if we change one SUDOKU rule so as to play a different game, we will have to train the AI model using another set of hundreds of thousands of solved puzzles of the new game. On the contrary, in the case of logic-based AI, we can have again 100% accuracy just by changing *one* symbolic representation, the symbolic representation of the rule that we altered (*ibid.*). To conclude, if an information-processing actor needs hundreds of thousands of data to perform the same task that another information-processing agent can perform by consulting a few lines of data and with higher accuracy and less sensitivity to changes, it can hardly be called *intelligent* the former intelligent. Note that in §III.3.3, we will see another notorious example of AI's stupidity which is premissed in its inability to represent meaningful information, the so-called *Clever Hans* problem.

Before concluding with the appropriate construal of IN-BETWEENNESS and subsequently of DISPLACEMENT 4.0, it should be noted that non-anthropomorphic AI can *not fully* diverge from human reason. It will always have components that are *symbolic* representations of human reason. As argued in §3.2.1.1, ¶5, every piece of technology has an *interface* that allows its interaction with the human user. For the user to use that interface, the latter needs to have a structure that the user can *semantically* interpret. In the case of AI, that interface is *minimally* the *input* & the *output* of the AI model.

interface's anthropomorphism

Summing up, a more accurate representation of the IN-BETWEENNESS is one where the human user initiates the use of technology to realise a personal non-technological need, and then, technology acquires a certain level of epistemic autonomy potentially misorienting the human user's command:

[84] A more precise definition of what constitutes the *subject matter* of a discipline will be given in §II.3.1.2.1, ¶3.

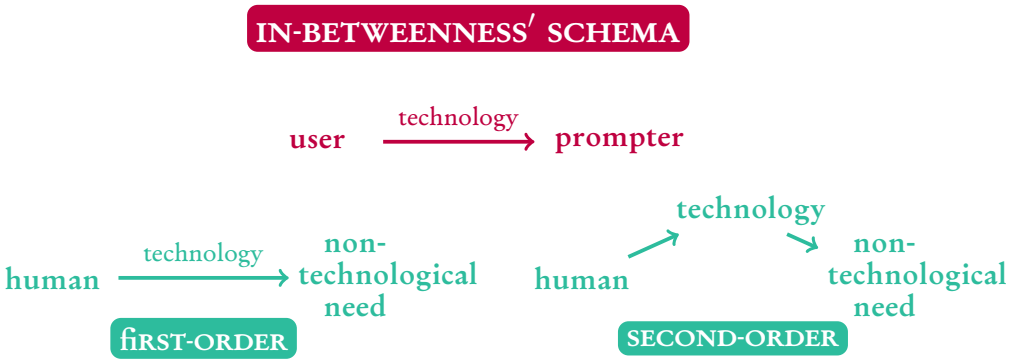


Figure 3: My reconstruction of IN-BETWEENNESS.

I.3.2.1.2 Ismene's dilemma

In Sophocles's "*Antigone*", Creon, king of Thebes, gives an order to leave the dead body of his nephew Polynices unburied outside the walls of Thebes to be eaten by the beasts as a punishment for waging a civil war. Polynices' sister Antigone decided to defy Creon's order and bury her brother's body. Doing otherwise would be against the sacred laws of gods. And for Antigone, no human laws can override the laws of the gods. For Creon's supporter's though, "*[t]he power is [Creon's],... to enforce it with the laws, both for the dead and all of us, the living*" (Sophocles (5th century BC) 1984, §Antigone) and hence, Creon does not hesitate to defy the laws of the gods and do as he deems right. He *rational*ises his choice by arguing that it is "*inconceivable*" for the gods to care about a traitor. Antigone ends up being arrested for her defiance of Creon's authority and Creon orders her death. A chorus of elder citizens enters the scene: "*Zeus, yours is the **power**, Zeus, **what man** on earth can override it, who can hold it back?*" (*ibid.*, emphasis added). Enlightenment responded that any "*man*" with *reason* can "*override*" Zeus' power. Even if there is a God-made law, the laws of a political order should be the laws that human reason dictates in its effort to approximate the laws of God (§2.3, ¶4).

Ismene's dilemma

More than 200 years after Enlightenment, ALGOAI brings us once more before the same question that Antigone, Creon, and Enlightenment's philosophes faced. It is the same *dilemma* that *Ismene*, Antigone's sister, had to confront. Should she honour the gods by helping her sister to bury her brother or should she honour the city by obeying the order of the king? Which of the two choices should she decide to *legitimise* her political act? Similarly, today, we have to answer which *ordo essendi* should constitute the *functional* dimension of an algocratic order. Should we accept a *misoriented ordo essendi* (co-)produced by humans & ALGOAI or should we reject any misorientation as *il*legitimate? I will call this dilemma *Ismene's dilemma*, the position of accepting a misoriented *ordo essendi* as the *post-Enlightenment horn* of the dilemma, and the position of rejecting a misoriented *ordo essendi* as the *Enlightenment horn*.

For the advocates of epistocracy the choice for Ismene's dilemma seems rather straightforward: ALGOAI can embody the ideal judge, the one we should obey in SOCIETY 5.0, and hence we should choose the post-Enlightenment horn. On the contrary, I advocate for the Enlightenment horn. To make my case, I first argue how the epistocratic position is defeated by Benacerraf's curse. Then, I argue that the value of the rule of law should still be an end of a post-Enlightenment order, a requirement that forces that post-Enlightenment order to return to its Enlightenment state.

the epistocratic argument

First things first, let's begin by presenting how the epistocratic argument supports the post-Enlightenment horn. Take for instance the example of the *Perincek v. Switzerland (2015)* case (§3.2.1, ¶3). According to epistocracy's *ontic tenet*, there exists a solution to the question of which are the properties of human rights that is independent of whether we include AI in the procedure of identifying that solution. This further entails that there exists an *ordo essendi* of values that grounds that solution that is independent of whether AI partakes in the procedure of acquiring knowledge about that *ordo essendi*. At the same time, according to the advocates of epistocracy, AI partaking in the procedure of identifying that solution raises the probability of identifying it. I.e., the *epistemic tenet* holds. Subsequently, according to the *authority tenet*, ALGOAI should be given the authority to co-produce power. At the same time, if the inclusion of human authorities in the procedure of identifying the solution lowers the probability of identifying it, then according to the *anti-authority tenet*, those human authorities should be deprived of their judicial authority and be superseded by *replacement* ALGOAI.

Benacerraf's curse *v.* epistocracy

The first problem with the epistocratic argument is Benacerraf's curse: we lack the paradigmatic knowledge that is required to be able to evaluate whether a procedure of identifying the solution in which ALGOAI partakes is more probable to identify the solution than the already existing procedure in which ALGOAI does not partake. Remember the example of AI used for medical diagnoses in §1.1.1, ¶6: if there is not enough consensus about

the already existing diagnoses, then we have no criterion standard to evaluate the accuracy of the AI model in the first place, let alone compare it to the accuracy of human medical experts. And in the case of ECtHR's judgements, the controversy among the experts is way more challenging than most of the controversies in empirical sciences. There are cases with tens of pages of dissenting opinions signed by multiple judges (e.g., in the *Perincek v. Switzerland (2015)* judgement, there are 11 pages of dissenting opinions signed by 12/17 judges and in the *Perincek v. Switzerland (2013)* judgement there are 22 pages of dissenting opinions signed by 4/7 judges). Subsequently, the epistocrat is left with two options. Either they have to admit that their *belief* that an AI-based judgement is in general more accurate is *not* rationally justified or they have to provide a convincing *rational* justification as to why Benacerraf's curse is lifted. If they do the former, then what we have is essentially a return to the *pre*-Enlightenment legitimacy paradigm of a *non*-rational *faith*. It is the non-rational faith that some entity has in general better epistemic capabilities from humans and ergo, their proposed solutions in general better even if humans lack the epistemic capabilities to rationally justify they are better. ALGOAI engineers end up being the clergy of our time that preaches on the advantages of those "*superior*" capabilities. For them AI can epistemically access "*ends for which there may be no category of human understanding*" (Kissinger 2018; *cf.* §3.2.1.1). In this case, the epistocrat then has to justify why we should return to what was abandon centuries ago, why in this *faith*-based system ALGOAI engineers differ from pre-Enlightenment's clerics, why ALGOAI's authority differs from pre-Enligthenment's divine right of kings (§2.3, ¶3).

The other option of the epistocrat is to attempt to *rationally* justify AI's superior epistemic capabilities, which is pretty much the option that the epistocratic philosophes of our time opt for. After all, the best way to argue why a paradigm has become obsolete is to show how another paradigm meets more adequately the conditions that justify the superseded paradigm. There can not be a bigger defeat for Enlightenment's legitimacy paradigm than showing *rationally* that it is not no longer enough. However, a rational justification of why Benacerraf's curse is lifted presupposes that there is sufficient paradigmatic knowledge in the first place that allows the comparison between the two types of knowledge acquisition procedures, those employed by human reason & those employed by non-anthropomorphic AI. Therefore, as long as human reason does not have epistemic access to such paradigmatic knowledge, there can not be an evaluation *via* human reason of which methodology is more adequate. Having said that, for many cases, the paradigmatic knowledge is sufficient enough to be able to perform such an evaluation. For instance, the ECtHR has decided to group cases that share common characteristics together, select one of those cases, make a final judgement about it (what it calls a *pilot judgement*), and then treat the rest of the cases similarly (ECtHR's Press Unit 2023cl *cf.* Dominik 2013). In those cases, we can take full advantage of the non-anthropomorphic AI's skills described in §3.1.

The epistocrat could counterargue that in the cases for which we lack paradigmatic knowledge, we are not only unable to evaluate algocratic authorities' output, but we are also unable to evaluate *human* authorities' output. Ergo, both outputs are *equally* dubious. One could even argue that in those cases there is no correct solution in the first place, and that the outcome can *not* be the result of a rational justification but of a personal *decision* performed by the judicial authority, a position that is called in the literature of jurisprudence *decisionism* and is attributed to philosopher Carl Schmitt (1888-1985) (Cristi 2014). Consequently, if a judgement is equally suboptimal regardless of whether it is made by a human or an ALGOAI authority, why refusing to delegate judicial authority to the latter? As we will see in §II.3.1.2.2, this is a common argument for those defending replacement legal ALGOAI: as long as legal ALGOAI is at most as problematic as human authorities, then we have no reason to reject it, especially if we consider its benefits on legitimacy that we saw in in §3.1 (faster justice, more accessible justice, etc).

the best of the worst

In this case, the debate boils down to the value of *self-governance*. If there are no rational criteria to determine what is the right solution, or at least the optimal solution, then, delegating the authority to decide how our political order should be is essentially an abandonment of the Enlightenment-rooted right of *self-governance* (§2.8, ¶3; §2.3, ¶8). A middle ground that allow us to retain our right to self-governance is to determine which are the cases where the interpretation & application of law becomes controversial enough to suffer from Benacerraf's curse. In those cases, human judges should be delegated with a more decisive role in the formation of the final judgement. However, even that solution is susceptible to misorientation, since as argued in §3.2.1, the convenience of using AI, the belief in AI's superior abilities, and other extraneous factors can bias human authorities to accept as non-controversial cases that they should have classified as controversial. Considering this, it is the proposal of the author that instead of deciding which cases are sensitive to Benacerraf's curse based on their *content*, we should make this decision based on the *reasoning methods* that judicial authorities would have used in those cases. The cases that suffer from Benacerraf's curse are those cases for which judicial reasoning meets a dead end, not being able to identify a final solution. By adopting this response, we do not escape Enlightenment's legitimacy paradigm since justifications are still grounded on human reason. Albeit, we can still use non-anthropomorphic AI that produces Chinese-room generated justifications for its output as long as those justifications reflect the appropriate type of judicial reasoning. If it looks like a judge, it *is* a judge!

self-governance

There is though another argument about retaining Enlightenment's legitimacy paradigm and use legal AlgoAI that justifies its output *via* human reason. And this time, it is not an argument to support a mere suggestion, but an argument to support an *obligation*. More precisely, the rule of law, at least its formal non-substantive requirements, should still be ends of the functional dimension of a post-Enlightenment political orders. As argued in §2.4, for a political order to be well-ordered, all of its actors need to be governed by the law (value of legality). At the same time, the actors need to have a sufficient level of epistemic access to those laws: they need to know how they are expected to act in order to act the way they are expected to act (values of legal certainty, open government, foreseeability, etc). For that to be possible, at least certain aspects of law 4.0 should be written in a way that allows *human reason* to *foresee* with sufficient *certainty* how law 4.0 will be applied as well as to *understand* law 4.0's application. At the same time, legal AlgoAI authorities should also be checked and balanced. The check and balances of judicial authorities that we saw in §2.4, ¶9 require *inter alia* to provide *just justifications* for the interpretation and application of law 4.0, justifications that should be epistemically accessible to everyone (written in plain language, translated, available *per* request, etc).

Take the example of predictive justice where the AlgoAI outputs whether there has been a violation of Convention's article 10 (freedom of expression) like the Court did in the *Perincek v. Switzerland (2015)* case. A mere binary result (violation/no violation) can not provide neither legal certainty nor foreseeability nor transparency nor sufficient for check-and-balancing information. There needs to be a *justification* about why the law was interpreted and applied the way it was interpreted and applied so as to decide the (non-)violation of article 10. In other words, the justification should *meaningfully* connect the *facts of a case* to the *law*, meaningfully to *us*, humans (*cf.* Iatrou 2022, §3; Adrien et al. 2021, §3.3.3; §II.4.1.2).

how justifications *should* be *legitimate* epistemic access

The question now is *which* are the *legitimate* methods to make such meaningful connections between the facts of the case and the law. According to the value of legality, it is *the law* itself that dictates how the interpretation & the application of the law should be performed. And the law designates specific judicial authorities with this task. Ergo, it is the reasoning methods that those authorities have developed in their case-law that are the legitimate methods of interpreting and applying the law. For instance, we will see in §IV.1, ¶6 that many times the US judicial authorities use in their criminal law practice a specific test to decide whether there is a causal relation between an act and a harm (e.g., the act of pulling a trigger and the harm of dying), the so-called *but-for test*. If an AlgoAI actor replaces those judicial authorities, then that AlgoAI should provide the *same* but-for justification in order to justify causal relations between acts and harm in criminal law cases.

a response to the objectivity challenge

Since we have determined which are the legitimate reasoning methods, we can now provide a first answer to the *objectivity challenge*. AlgoAI engineers should follow the construal of values and the methods of construing those values found in the practice of the authorities which are legitimately designated with the task of *interpreting* the political order's functional dimension: judicial authorities. In other words, what the designated judicial authority judges according to the procedures prescribed by the law *is* what is the case. In contrast to other disciplines, experts in law have the authority to establish their own *ordo essendi* (Alchourrón 2015).

Having said that, judicial authorities like the ECtHR are notoriously *inexact* in the reasoning methods they use to interpret and apply the law (*see e.g.* Letwin 2021; Mchangama and Alkiviadou 2021). For instance, the conditions under which the Court judges whether there exists a causal relation between the acts of the defendant and the alleged harm as well as whether the defendant is responsible for that harm are many times *contradictory* with each other, *underdetermined*, too *general*, etc (Letwin 2021; Stoyanova 2018 *cf.* Lavrysen 2018; §IV.1.1).[85] Differences in the reasoning methods of interpreting and applying the law are *decisive* the *ordo essendi* produced by judicial authorities. E.g., there can be conceptualisation of causal inference according to which a defendant is legally responsible for the alledged harm while according to other conceptualisations of causal inference, the same defendant is *not* legally responsible (*ibid.*). Ergo, we have two different conceptualisations of the value of legal responsibility and subsequently two distinct ***incompatible*** *ordo essendi*. This undermines many of the rule of law legitimacy requirements laid out in §2.4. It undermines legal certainty since there are contradictions and ambiguity in the case-law. It further undermines foreseeability since the subjects can not foresee how to act so as to abide by the law. It further underines check-and-balances since the judgments are justified with incompatible justifications. It also violates the value of legality since the law is not applied to everyone as it *should*. E.g., if according to the correct *ordo essendi* the defendant is legally responsible for the harm but judicial authorities judge that they are not responsible then the defendant has been allowed to act *unlawfully* without any consequences. The harm to legitimacy is exacerbated even further if one adds to the foregoing the human right requirements for a fair trial (*see e.g.*, the requirements listed in article 6, ¶3 in the §appendix).

accommodating inexactness

Considering the above, one could argue that AlgoAI engineers should correct cases of inexactness in judicial reaosning when they engineer AlgoAI models allowing the latter to surpass human judges' capabilities. Albeit counterintuitive, such a decision would end up engineering an *illegitimate* model. Firstly, as we will see in

[85] In §III.3.2.2, I explain in more detail different types of *inexactness* in judicial reasoning like types of logical contradictions, generality, underdetermination, overdetermination, etc.

§III.3.2.2, many cases of inexactness are actually desirable in judicial reasoning; trying to resolve them leads in reality to suboptimal judgements. But even if a case of inexactness is indeed a deficit in judicial reaonsing, ALGOAI engineers deciding how to resolve it is still an *illegitimate* decision. It is *judicial authorities* and not epistemic authorities which are prescribed by law with the task to interpret and apply the law.[86] Ergo, any changes to the interpretation & application of the law should be performed by judicial authorities. Otherwise, we have a violation of the value of legality. That being said, we can always introduce new LAW 2.0 that will allow ALGOAI engineers to signal back to judicial authorities such cases of inexcatenss leaving it up to them to decide how to resolve them. This is nothing more than the rule of law requirement of *institutionalising checks and balances* to prevent misuse/abuse of the interpretation & application of the law. more checks and balances

I.3.3 Conclusion: the necessity of logicians

Summing up, legal ALGOAI that contributes to the interpretation & application of the law needs to provide a ***rational** justification* of why the law should be interpreted & applied according to its output. This is why *logic & formal philosophy* should be irreplaceable parts of ALGOAI engineering. More precisely, both disciplines are concerned with the *identification & formalisation* of rational reasoning methods including judicial reasoning (*see* §II.4.1.2, *fn.* 30 for a detailed list of citations from logic & formal philosophy on the different types of rational judicial reasoning; *see* §III.2.1, ¶2 for a more detailed account of what I construe as *formal* philosophy). Those formalisations of rational judicial reasoning should be incorporated in the ALGOAI model either as a type of anthropomorphic AI or as a Chinese-room imitation of human reaosning or as formal restrictions to state-of-the-art connectionist AI that force it to have anthropomorphic components. In §II.4.2.3, I explain how can such hybrid AI models be engineered and why should legal ALGOAI engineering go to that direction. A position that is raising in popularity due to LAW 2.0 requirements of *justifying* the output of ALGOAI (§II.4.2.2, ¶¶3-4). Furthermore, it is logicians & formal philosophers that have the expertise to identify cases of reaosning inexactness in judicial arguments, and ergo, it is them that can have the most substantial contribution to check and balancing judicial authorities about their rationale, both human and algocratic judicial authorities. They can also guide the public to the *understand* how the justifications provided by ALGOAI can be used to *comprehend & foresee* the application of the law. In other words, logicians & formal philosophers should have a central role in ALGOAI engineering not only as ALGOAI engineers, but more importantly in terms of *legitimacy*, as *epistemic authorities* that check and balance other judicial and epistemic authorities, as well as *philosophes* that shape the pubic discourse about what how a legitimate algocratic SOCIETY 5.0 *should* be. It is their expertise that can *rationalise* ALGOAI's behaviour saving Enlightenment from its foretold death.

References

Švedkauskas, Žilvinas. 2022. "Digital surveillance, master key for MENA autocrats." In *Liberty doom? Artificial intelligence in Middle Eastern security,* edited by Justine Leïla Belaïd, 36–49. 27. Published by the European Institute of the Mediterranean, May.

Abraham, Tara H. 2002. "(Physio)logical circuits: the intellectual origins of the McCulloch-Pitts neural networks." *Journal of the history of the behavioral sciences* 38 (1): 3–25. https://doi.org/10.1002/jhbs.1094.

Abu-Taieh, Evon, Issam H. Al Hadid, and Ali Zolait. 2020. "5G road map to communication revolution." Chap. 1 in *Cyberspace,* edited by Evon Abu-Taieh, Abdelkrim El Mouatasim, and Issam H. Al Hadidr, Section 1: Internet and communications, 3–12. IntechOpen. https://doi.org/10.5772/intechopen.7888.

Adorno, Theodor W., and Max Horkheimer. (1947) 2002. *Dialectic of Enlightenment.* Edited by Gunzelin Schmid Noerr. Translated by Edmund Jephcott. Stanford University Press.

Adriaans, Pieter. 2020. "Information." In *The Stanford Encyclopedia of Philosophy,* Fall 2020, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Adrien, Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. 2021. "Legal requirements on explainability in machine learning." *Artificial Intelligence and Law* 29:149–169. https://doi.org/10.1007/s10506-020-09270-4.

AFP (Agence France-Presse). 2023. *China's Xi Jinping calls for greater state control of AI to counter "dangerous storms": President says national security threats are increasing and urged greater oversight of artificial intelligence*

[86] This is another instantiation of *Ismene's dilemma*: should we accept the faulty *ordo essendi* produced by the legitimate authorities or should we reject it as unjust?

and data security. China: The Guardian, June 8, 2023. Accessed June 1, 2023. https://www.theguardian.com/world/2023/jun/01/chinas-xi-jinping-calls-for-greater-state-control-of-ai-to-counter-dangerous-storms.

Agar, Jon. 2020. "What is science for? The Lighthill report on artificial intelligence reinterpreted." *The British Journal for the History of Science* 53 (3): 289–310. https://doi.org/10.1017/S0007087420000230.

Akhtar, Rais, ed. 2022. *Coronavirus (COVID-19) outbreaks, vaccination, politics and society: The continuing challenge.* Springer Cham. https://doi.org/10.1007/978-3-031-09432-3.

Aksoy, Pınar Cağlayan. 2022. "AI as agents: Agency law." Chap. 11 in *The Cambridge handbook of artificial intelligence: Global perspectives on law and ethics,* edited by Larry A. DiMatteo, Cristina Poncibò, and Michel Cannarsa, Part III: AI and liability, 146–162. Cambridge University Press.

Alchourrón, Carlos E. 2015. "Limits of logic and legal reasoning." In *Essays in legal philosophy,* Reprint, edited by Carlos Bernal and Carla Huerta. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198729365.003.0017.

Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. 2016. "Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective." *PeerJ Computer Science* 2:e93.

Alexander, Larry, and Emily Sherwin. 2008. *Demystifying legal reasoning.* Cambridge Introductions to Philosophy and Law series. Cambridge University Press.

Ali, Arden, Sally Haslanger, Jerome Hodges, and Lily Hu. n.d. *Encoding race and structural racism: philosophical perspectives.* Work in progress. Accessed June 12, 2023. https://sallyhaslanger.weebly.com/uploads/1/8/2/7/18272031/ali_haslanger_hodges_hu__encoding_race_and_structural_racism_circulating_draft_may_2021_updated_.pdf.

Allen, Amy. 2022. "Feminist perspectives on power." In *The Stanford Encyclopedia of Philosophy,* Fall 2022, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Amir, Eyal. 2014. "Reasoning and decision making." Chap. 9 in *The Cambridge handbook of artificial intelligence,* edited by Keith Frankish and William M. Ramsey, Part III: Dimensions, 191–212. Cambridge University Press.

Anderson, Jackie, and Kenn R. Ghaffarian. 2023. *Early pregnancy diagnosis.* Online. StatPearls Publishing, January 2, 2023. https://www.ncbi.nlm.nih.gov/books/NBK556135/.

Araujo, Leone. 2022. "Pivotal legal and ethical caveats for artificial intelligence in the Middle East." *Journal of Law in the Middle East* 2.

Arnardóttir, Oddný Mjöll. 2016. "Rethinking the two margins of appreciation." *European Constitutional Law Review* 12 (1): 27–53. https://doi.org/10.1017/S1574019616000018.

Arvidsson-Shukur, D. R. M., A. N. O. Gottfries, and C. H. W. Barnes. 2017. "Evaluation of counterfactuality in counterfactual communication protocols." *Physical Review A* 96 (062316).

Austin, John. (1832) 1955. *The province of jurisprudence determined.* Edited by H. L. A. Hart. Weidenfeld & Nickolson.

Bai, Chunguang, Patrick Dallasega, Guido Orzes, and Joseph Sarkis. 2020. "Industry 4.0 technologies assessment: A sustainability perspective." *International Journal of Production Economics* 229:107776. https://doi.org/10.1016/j.ijpe.2020.107776.

Baker-Beall, Christopher, and Gareth Mott. 2021. "Understanding the European Union's perception of the threat of cyberterrorism: A discursive analysis." *JCMS: Journal of Common Market Studies* 60 (4): 1086–1105. https://doi.org/10.1111/jcms.13300.

Barma, Naazneen H. 2016. *The peacebuilding puzzle: Political order in post-conflict states.* Cambridge University Press. https://doi.org/10.1017/9781316718513.

Barrett, Paul. 2020. *Who moderates the social media giants? A call to end outsourcing.* Report. New York University (NYU) Stern Center for Business and Human Rights, June 8, 2020.

Bayer, Judit, and Petra Bárd. 2020. *Hate speech and hate crime in the EU and the evaluation of online content regulation approaches.* Study requested by the LIBE Committee. European Parliament Think Tank. Accessed October 27, 2021. https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2020)655135.

Beatson, Jack. 2021. *The rule of law and the separation of powers.* Key ideas in law series. Hart Publishing.

Beetham, David. 2013. *The legitimisation of power.* 2nd ed. Edited by B. Guy Peters, Jon Pierre, and Gerry Stoker. Series: Political Analysis. Palgrave Macmillan.

Benacerraf, Paul. 1983. "Mathematical truth." In *Philosophy of mathematics: Selected readings*, 2nd ed., edited by Paul Benacerraf and Hilary Putnam. Cambridge University Press.

Berlin, Isaiah. 1993. *The Magus of the North: J. G. Hamann and the origins of modern irrationalism.* Edited by Henry Hardy. John Murray.

Binchy, Emily. 2022. *Advancement or impediment? AI and the rule of law.* Published by the Institute of International and European Affairs (IIEA). https://www.iiea.com/publications/advancement-or-impediment-ai-and-the-rule-of-law.

Blau, Adrian. 2023. "Political equality and political sufficiency." *Moral Philosophy and Politics* 10 (1): 23–46. https://doi.org/10.1515/mopp-2020-0059.

Boden, Margaret A. 2014. "GOFAI." Chap. 4 in *The Cambridge handbook of artificial intelligence*, edited by Keith Frankish and William M. Ramsey, Part II: Architecture, 89–107. Cambridge University Press.

Bongiovanni, Giorgio, Gerald Postema, Antonino Rotolo, Giovanni Sartor, Chiara Valentini, and Douglas Walton, eds. 2018. *Handbook of legal reasoning and argumentation.* Springer, Dordrecht.

Brennan, Jason. 2016. *Against democracy.* Princeton University Press.

Bristow, William. 2017. "Enlightenment." In *The Stanford Encyclopedia of Philosophy*, Fall 2017, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Britannica (Editors of Encyclopaedia). 2021. *Newton's laws of motion.* Encyclopedia Britannica, March 18, 2021. Accessed April 10, 2023. https://www.britannica.com/science/Newtons-laws-of-motion.

———. 2023. *Divine right of kings.* Encyclopedia Britannica, March 27, 2023. Accessed April 10, 2023. https://www.britannica.com/topic/divine-right-of-kings.

Brown, Garrett W., Iain McLean, and Alistair McMillan, eds. 2018. *A concise Oxford dictionary of politics and international relations.* 4th ed. Oxford University Press.

Brownsword, Roger. 2021. *Law 3.0: Rules, regulation, and technology.* Law/science and technology studies. Routledge.

———. 2022. "Law, authority, and respect: three waves of technological disruption." *Law, Innovation and Technology* 14 (1): 5–40. https://doi.org/10.1080/17579961.2022.2047517.

Buckner, Cameron, and James Garson. 2019. "Connectionism." In *The Stanford Encyclopedia of Philosophy*, Fall 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Burrows, Simon. 2015. "Books, Philosophy, Enlightenment." Chap. 5 in *The Oxford Handbook of the French Revolution*, edited by David Andress, Part I: Origins, 74–91. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199639748.013.005.

Buss, Sarah, and Andrea Westlund. 2018. "Personal autonomy." In *The Stanford Encyclopedia of Philosophy*, Spring 2018, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Bylund, Per L., and Mark D. Packard. 2021. "Separation of power and expertise: Evidence of the tyranny of experts in Sweden's COVID-19 responses." *Southern Economic Journal* 87 (4): 1300–1319. https://doi.org/https://doi.org/10.1002/soej.12493.

Byrne, David, and Carol Corrado. 2016. *ICT asset prices: Marshaling evidence into new measures.* Economics Program Working Paper Series (EPWP) #16-06. The Conference Board, July. https://www.conference-board.org/publications/ICT-Asset-Prices-Evidence-to-New-Measures.

Cafaro, Susanna. 2023. "The postwar European integration process and the progressive construction of a supranational legal order." Chap. 4 in *How democracy survives: Global challenges in the Anthropocene*, edited by Michael Holm and R. S. Deese, Part I: The forgotten promise of 1945, 65–80. Democratization and Autocratization Studies. Routledge.

Cahoone, Lawrence. 2023. "The end of Enlightenment liberalism?" *The Journal of Speculative Philosophy* 37 (1): 81–98.

Campbell, David F. J., and Wieland Schneider. 2020. "Media and innovation." In *Encyclopedia of creativity, invention, innovation and entrepreneurship*, 2nd ed., edited by Elias G. Carayannis. Springer Cham.

Cantù, Paola. 2014. "The right order of concepts: Graßmann, Peano, Gödel and the inheritance of Leibniz's universal characteristic." *Philosophia Scientiæ* 18-1 (1): 157–182. 10.4000/philosophiascientiae.921.

Capaldi, Nicholas. 1975. *David Hume: The Newtonian philosopher.* Twayne.

———. 1998. *The Enlightenment project in the analytic conversation.* Edited by H. Tristram Engelhardt. Vol. 4. Philosophical Studies in Contemporary Culture. Springer Netherlands. https://doi.org/10.1007/978-94-017-3300-7_1.

Carayannis, Elias G., and Joanna Morawska-Jancelewicz. 2022. "The futures of Europe: Society 5.0 and Industry 5.0 as driving forces of future universities." 13 (4): 3445–3471. https://doi.org/0.1007/s13132-021-00854-2.

Carnap, Rudolf. (1928) 1967. *The logical structure of the world and pseudoproblems in philosophy (orig: Der logische Aufbau der Welt (aka Aufbau)).* Translated by Rolf A. George. University of California Press.

———. 1938. *The logical syntax of language.* Harcourt, Brace and Company.

CEPEJ (European Commission for the Efficiency of Justice). 2019a. "Appendix I: In-depth study on the use of AI in judicial systems, notably AI applications processing judicial decisions and data." In *European ethical Charter on the use of artificial intelligence in judicial systems and their environment.* Prepared by Xavier Ronsin and Vasileios Lampos. Printed by the Council of Europe. https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment.

———. 2019b. *European ethical Charter on the use of artificial intelligence in judicial systems and their environment.* Printed by the Council of Europe. https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment.

Chalkidis, Ilias, Ion Androutsopoulos, and Nikolaos Aletras. 2019. "Neural legal judgment prediction in English." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4317–4323. Florence, Italy: Association for Computational Linguistics, July. https://doi.org/10.18653/v1/P19-1424.

Chang, Hasok. 2021. "Operationalism." In *The Stanford Encyclopedia of Philosophy*, Fall 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Chatziathanasiou, Konstantin. 2022. "Beware the lure of narratives: "hungry judges" should not motivate the use of "artificial intelligence" in law." *German Law Journal* 23 (4): 452–464. https://doi.org/10.1017/glj.2022.32.

Chen, Rong. 2021. "Causal network inference for neural ensemble activity." *Neuroinformatics* 19 (3): 515–527. https://doi.org/10.1007/s12021-020-09505-4.

Chevalier-Watts, Juliet. 2010. "Has human rights law become lex specialis for the European Court of Human Rights in right to life cases arising from internal armed conflicts?" *The International Journal of Human Rights* 14 (4): 584–602. https://doi.org/10.1080/13642980903205383.

Chomanski, Bartek. 2022. "Legitimacy and automated decisions: the moral limits of algocracy." *Ethics and information technology* 24 (34). https://doi.org/10.1007/s10676-022-09647-w.

Christensen, Tom, and Per Lægreid. 2020. "Balancing governance capacity and legitimacy: How the Norwegian government handled the COVID-19 crisis as a high performer." *Public Administration Review* 80 (5): 774–779. https://doi.org/https://doi.org/10.1111/puar.13241.

Christiano, Thomas. 2020. "Authority." In *The Stanford Encyclopedia of Philosophy*, Summer 2020, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Christina, Schori L. 2017. "Unveiling the "United Cyber Caliphate" and the birth of the e-terrorist." *Georgetown Journal of International Affairs* 18 (3): 11–20.

Chu, Yun-Han. 2016. "Sources of regime legitimacy in Confucian societies." *Journal of Chinese Governance* 1 (2): 195–213. https://doi.org/10.1080/23812346.2016.1172402.

Clark, Andy. 1990. "Connectionism, competence, and explanation." Chap. 12 in *The philosophy of artificial intelligence*, edited by Margaret A. Boden, 281–308. Oxford readings in philosophy. Oxford University Press.

Cohon, Rachel. 2018. "Hume's Moral Philosophy." In *The Stanford Encyclopedia of Philosophy*, Fall 2018, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Colbourn, Trevor. 1998. *The lamp of experience: Whig history and the intellectual origins of the American Revolution*. Reprint. Liberty Fund.

Cole, David. 2023. "The Chinese room argument." In *The Stanford Encyclopedia of Philosophy*, Summer 2023, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Commissioner for Human Rights. 2019. *Unboxing artificial intelligence: 10 steps to protect Human Rights*. Printed at the Council of Europe, May.

Copernicus, Nicolaus. (1543) 1995. *On the revolutions of heavenly spheres*. Translated by Charles Glenn Wallis. Great Minds Series. Prometheus.

Cristi, Renato. 2014. "Decisionism." In *The Encyclopedia of Political Thought*, 831–833. John Wiley Sons, Ltd. https://doi.org/10.1002/9781118474396.wbept0244.

Crook, Malcolm. 2015. "The new regime: Political institutions and democratic practices under the constitutional monarchy, 1789–91." Chap. 13 in *The Oxford Handbook of the French Revolution*, edited by David Andress, Part III: Revolution and constitution, 218–235. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199639748.013.013.

D'Agostino, Fred, Gerald Gaus, and John Thrasher. 2021. "Contemporary Approaches to the Social Contract." In *The Stanford Encyclopedia of Philosophy*, Winter 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Danaher, John. 2016. "The threat of algocracy: Reality, resistance and accommodation." *Philosophy and Technology* 29 (3): 245–268. https://doi.org/10.1007/s13347-015-0211-1.

Danks, David. 2014. "Learning." Chap. 4 in *The Cambridge handbook of artificial intelligence: Responsible artificial intelligence*, edited by Keith Frankish and William M. Ramsey, Part III: Dimensions. Cambridge University Press.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous factors in judicial decisions." *Proceedings of the National Academy of Sciences* 108 (17): 6889–6892. https://doi.org/10.1073/pnas.1018033108.

de Jong, Willem R., and Arianna Betti. 2010. "The classical model of science: a millennia-old model of scientific rationality." *Synthese* 174 (2): 185–203. https://doi.org/10.1007/s11229-008-9417-4.

de la Mettrie, Julien Offray. (1747) 1912. *Man a machine (orig: L'homme machine): Including Frederick the Great's "Eulogy" on la Mettrie and extracts from la Mettrie's "The natural history of the soul"*. Philosophical and historical notes by Gertude Carman Bussey. The Open Court Publishing Co. https://www.gutenberg.org/files/52090/52090-h/52090-h.htm.

DeLaney, Matthew, and Luke Wood. 2021. "Urine pregnancy testing: When does no mean maybe?" *Journal of the American College of Emergency Physicians Open* 2 (5): e12567. https://doi.org/10.1002/emp2.12567.

Devetak, Richard, and Tim Dunne. 2005. "Order." In *Encyclopedia of international relations and global politics*, edited by Griffiths Martin, 613–622. Routledge.

Dinniss, Heather A. Harrison. 2018. "The threat of cyber terrorism and what international law should (try to) do about it." *Georgetown Journal of International Affairs* 19:43–50.

Dirks, Michael, Nicolas K. Ewerbeck, Tobias M. Ballhause, Sebastian Weiß, Andreas Luebke, Carsten Schlickewei, Karl-Heinz Frosch, and Matthias Priemel. 2023. "The diagnostic accuracy of 332 incisional biopsies

in patients with malignant tumors in the musculoskeletal system." *World Journal of Surgical Oncology* 21 (1). https://doi.org/10.1186/s12957-022-02883-w.

Doh, Yonsoo. 2020. "Why the South Korean public wants AI judges: How AI can and could complement human's decision making" (October 31, 2020). Accessed March 13, 2023. https://medium.com/carre4/why-the-south-korean-public-wants-ai-judges-ed8f52a4e573.

Dominik, Haider. 2013. *The pilot-judgment procedure of the European Court of Human Rights.* Brill | Nijhoff.

Donn, Natasha. 2023. *Artificial intelligence can identify risks in public contracting – court.* Portugal Resident, January 30, 2023. Accessed March 10, 2023. https://www.portugalresident.com/artificial-intelligence-can-identify-risks-in-public-contracting-court/.

Downey, Cameron. 2021. "Roger Brownsword (2020) Law 3.0: Rules, regulation and technology. Abingdon: Routledge." *Law, Technology and Humans* 3 (1): 151–153. https://doi.org/10.5204/lthj.1838.

Duignan, Brian. 2023. *Enlightenment.* Encyclopedia Britannica, May 15, 2023. Accessed May 20, 2023. https://www.britannica.com/event/Enlightenment-European-history.

Dworkin, Ronald. 1986. *Law's empire.* Harvard University Press.

———. 2011. *Justice for hedgehogs.* Belknap Press.

Dzehtsiarou, Kanstantsin. 2011. "European consensus and the evolutive interpretation of the European Convention on Human Rights." *German Law Journal* 12 (10): 1730–1745. https://doi.org/10.1017/S2071832200017533.

———. 2015. *European consensus and the legitimacy of the European Court of Human Rights.* Cambridge University Press. https://doi.org/10.1017/CBO9781139644471.

EC (European Commission). 2023. *New tax transparency rules will help Member States shine a light on the crypto-asset sector.* Press Release. Brussels, May 16, 2023. Accessed May 25, 2023. https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2725.

ECtHR Registry. 2021. *Guide on Article 10 of the European Convention on Human Rights: Freedom of expression.* Updated. April. https://www.echr.coe.int/documents/guide_art_10_eng.pdf.

ECtHR's Press Unit (Unité de la Presse). 2023a. *Factsheet on climate change.* March. Accessed April 28, 2023. https://www.echr.coe.int/Pages/home.aspx?p=press/factsheets&c.

———. 2023b. *Factsheet on new technologies.* January. Accessed April 28, 2023. https://www.echr.coe.int/Pages/home.aspx?p=press/factsheets&c.

———. 2023c. *Factsheet on pilot judgements.* March. Accessed April 28, 2023.

———. 2023d. *Factsheet on work-related rights.* February. Accessed April 28, 2023.

ECtHR's Registry. 2022a. *Guide on Article 1 of the European Convention on Human Rights: Obligation to respect human rights – Concepts of "jurisdiction" and imputability.* Updated. August. https://www.echr.coe.int/documents/guide_art_1_eng.pdf.

———. 2022b. *Guide on Article 6 of the European Convention on Human Rights: Right to a fair trial (criminal limb).* Updated. August. https://www.echr.coe.int/documents/guide_art_6_eng.pdf.

———. 2022c. *Guide on Article 6 of the European Convention on Human Rights: Right to respect for private and family life, home and correspondence.* Updated. August. https://www.echr.coe.int/documents/guide_art_6_criminal_eng.pdf.

———. 2022d. *Guide on Article 8 of the European Convention on Human Rights: Right to respect for private and family life, home and correspondence.* Updated. August. https://www.echr.coe.int/documents/guide_art_8_eng.pdf.

Edelstein, Melvin. 2014. *The French Revolution and the birth of democracy.* Ashgate.

Ehrenberg, Kenneth M. 2008. "Archimedean metaethics defended." *Metaphilosophy* 39 (4/5): 508–529.

EIU (Economist Intelligence Unit). 2023. *Democracy Index 2022: Frontline democracy and the battle for Ukraine.* Published by the Economist Intelligence Unit (EIU). https://www.eiu.com/n/campaigns/democracy-index-2022/.

Eren, Ozkan, and Naci Mocan. 2018. "Emotional judges and unlucky juveniles†." *American Economic Journal: Applied Economics* 10 (3): 171–205. https://doi.org/10.1257/app.20160390.

Estlund, David. 1993. "Making truth safe for democracy." In *The idea of democracy*, edited by David Copp, Jean Hampton, and John E. Roemer, 53–68. Cambridge University Press.

———. 2003. "Why not Epistocracy?" In *Desire, identity, and existence: Essays in honour of T. M. Penner*, edited by Naomi Reshotko, 53–68. Academic Printing / Publishing.

———. 2008. *Democratic authority: A philosophical framework*. Princeton University Press.

Etienne, Hubert. 2021. "The dark side of the 'Moral Machine' and the fallacy of computational ethical decision-making for autonomous vehicles." *Law, Innovation and Technology* 13 (1): 85–107. https://doi.org/10.1080/17579961.2021.1898310.

Etzioni, Amitai. 2010. "The normativity of human rights is self-evident." *Human Rights Quarterly* 32 (1): 187–197.

eu-LISA and Eurojust. 2022. *Artificial intelligence supporting cross-border cooperation in criminal justice*. Joint report. June. https://doi.org/0.2857/364146.

Fagan, Andrew. n.d. *Human Rights*. Internet Encyclopedia of Philosophy. Accessed May 5, 2023. https://iep.utm.edu/hum-rts/.

Fan, Yaxiang, Min Zheng, Bo Liu, and Le Sun. 2022. "LawRec: Automatic recommendation of legal provisions based on legal text analysis." *Computational Intelligence and Neuroscience*, 6313161. https://doi.org/10.1155/2022/6313161.

Fenwick, Mark, and Stefan Wrbka. 2016. "The shifting meaning of legal certainty." In *Legal certainty in a contemporary context: Private and criminal law perspectives*, edited by Mark Fenwick and Stefan Wrbka, 1–6. Springer Singapore. https://doi.org/10.1007/978-981-10-0114-7_1.

Fitzsimmons, Michael P. 2015. "Sovereignty and constitutional power." Chap. 12 in *The Oxford Handbook of the French Revolution*, edited by David Andress, Part III: Revolution and constitution, 201–217. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199639748.013.012.

Floridi, Luciano. 2014. *The 4th Revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Folarin, Sheriff, Esther Akinlabi, and Aderemi Atayero, eds. 2022. "The United Nations and Sustainable Sevelopment Goals," https://doi.org/10.1007/978-3-030-95971-5.

Foner, Eric. 2005. "Tom Paine and revolutionary America."

Fordham, Michael. 2020. *Judicial review handbook*. 7th ed. Hart Publishing.

Franklin, Stan. 2014. "History, motivations, and core themes." Chap. 1 in *The Cambridge handbook of artificial intelligence*, edited by Keith Frankish and William M. Ramsey, Part I: Foundations, 15–33. Cambridge University Press.

Fritsch, Lothar, Aws Jaber, and Anis Yazidi. 2022. "An overview of artificial intelligence used in malware." In *Nordic Artificial Intelligence Research and Development*, edited by Evi Zouganeli, Anis Yazidi, Gustavo Mello, and Pedro Lind, 41–51. Springer International Publishing.

Fromm, Christian, Antonios Likourezos, Lawrence Haines, Abu N. G. A. Khan, Janet Williams, and Joel Berezow. 2012. "Substituting whole blood for urine in a bedside pregnancy test." *The Journal of emergency medicine* 43 (3): 478–482. https://doi.org/10.1016/j.jemermed.2011.05.028.

Fukuyama, Francis. 2011. *The origins of political order: From prehuman times to the French Revolution*. Farrar, Straus / Giroux.

———. 2014. *Political order and political decay: From the industrial revolution to the globalization of democracy*. Farrar, Straus / Giroux.

Fukuyama, Mayumi. 2018. "Society 5.0: Aiming for a new human-centered society." *Japan Spotlight* 1:47–50.

Fung, Pascale, and Hubert Etienne. 2022. "Confucius, cyberpunk and Mr. Science: comparing AI ethics principles between China and the EU." *AI and Ethics*, https://doi.org/10.1007/s43681-022-00180-6.

Gálik, Slavomír, and Sabína Gáliková Tolnaiová. 2020. "Cyberspace as a new existential dimension of man." Chap. 2 in *Cyberspace*, edited by Evon Abu-Taieh, Abdelkrim El Mouatasim, and Issam H. Al Hadidr, Section 1: Internet and communications, 13–26. IntechOpen. https://doi.org/10.5772/intechopen.7888.

Gaudemet, Yves. 2015. "L'ordre public: Propos introductifs." *Archives de philosophie du droit* 58 (1): 3–4. https://doi.org/10.3917/apd.581.0028.

Gibbs, Samuel. 2016. *Chatbot lawyer overturns 160,000 parking tickets in London and New York: Free service DoNotPay helps appeal over $4m in parking fines in just 21 months, but is just the tip of the legal AI iceberg for its 19-year-old creator.* Chatbots. The Guardian, June 28, 2016. Accessed July 10, 2023. https://www.theguardian.com/technology/2016/jun/28/chatbot-ai-lawyer-donotpay-parking-tickets-london-new-york.

Gibson, William. 1984. *Neuromancerr.* Ace.

Goodman, Biyce, and Seth Flaxman. 2017. "European Union regulations on algorithmic decision-making and a "right to explanation"." *AI Magazine* 38 (3).

Goodwin, Jazmin. 2020. *Google AI system can surpass human experts in spotting breast cancer, study finds.* USA Today, January 3, 2020. Accessed January 10, 2023. https://eu.usatoday.com/story/tech/2020/01/03/google-ai-system-can-beat-human-experts-spotting-breast-cancer/2795100001/.

Governatori, Guido, Antonino Rotolo, and Giovanni Sartor. 2021. "Logic and the law: philosophical foundations, deontics, and defeasible reasoning." Chap. 9 in *Handbook of deontic logic and normative systems*, edited by Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, vol. 2. College Publications.

Gray, John. 1995. *Enlightenment's wake.* Routledge Classics. Routledge.

Gurney, Kevin. 2004. *An introduction to neural networks.* 1st ed. CRC Press.

Gutwirth, Serge, Ronald Leenes, Paul de Hert, and Yves Poullet, eds. 2013. *European data protection: Coming of age.* Springer Dordrecht. https://doi.org/10.1007/978-94-007-5170-5.

Habermas, Jürgen. 1992. *Legitimation crisis.* Reprint. Translated by Thomas McCarthy. *orig: Legitimitätsprobleme im spätkapitalismus*. Polity Press.

Hage, Jaap. 2005. *Studies in Legal Logic.* Vol. 70. Law and Philosophy Library. Springer Dordrecht. https://doi.org/10.1007/1-4020-3552-7.

Hale, Bob, and Crispin Wright. 2002. "Benacerraf's dilemma revisited." *European Journal of Philosophy* 10 (1): 101–129. https://doi.org/https://doi.org/10.1111/1468-0378.00151.

Hart, H. L. A. 1961. *The concept of law.* Clarendon Law Series. Oxford University Press.

Haynes, Jeffrey. 2019. "Introduction: The "Clash of Civilizations" and relations between the West and the Muslim world." *The Review of Faith & International Affairs* 17 (1): 1–10. https://doi.org/10.1080/15570274.2019.1570756.

Hebb, Donald O. 1949. *The organization of behavior: A neuropsychological theory.* Wiley.

Heldt, Amélie Pia. 2019. "Upload-filters: bypassing classical concepts of censorship?" *Journal of Intellectual Property, Information Technology, and Electronic Commerce Law (JIPITEC)* 10 (1): 56–64. http://nbn-resolving.de/urn:nbn:de:0009-29-4877.

Heller, Nathan. 2016. *Imagining a cashless world: Sweden shows us what life without paper currency might be like.* Stockholm: The New Yorker, October 3, 2016. Accessed May 15, 2023. https://www.newyorker.com/magazine/2016/10/10/imagining-a-cashless-world.

Hill, Charles. 2010. *Grand strategies: Literature, statecraft, and world order.* Yale University Press.

Hilpinen, Risto, and Paul McNamara. 2021. "Deontic logic: A historical survey and introduction." In *Handbook of deontic logic and normative systems*, edited by Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, vol. 1, Part I: Background, 3–136. College Publications.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. 2006. "A fast learning algorithm for deep belief nets." *Neural Computation* 18 (7): 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527.

Hobbes, Thomas. (1651) 2017. *Leviathan.* Edited by Christopher Brooke. Penguin Classics. Penguin Books.

Holm, Michael. 2023. "The other American Dream: The one world order and Human Rights." Chap. 1 in *How democracy survives: Global challenges in the Anthropocene,* edited by Michael Holm and R. S. Deese, Part I: The forgotten promise of 1945, 9–28. Democratization and Autocratization Studies. Routledge.

Horowitz, Irving Louis. 2006. *The struggle for democracy: The promotion of democracy is the centerpiece of Bush's foreign policy, but the president has yet to define democracy.* The National Interest, March 1, 2006. Accessed April 11, 2023. https://nationalinterest.org/article/the-struggle-for-democracy-880.

Hume, David. (1739) 2022. *A treatise of human nature.* Project Gutenberg, November 24, 2022. https://www.gutenberg.org/files/4705/4705-h/4705-h.htm#link2H_4_0086.

Huntington, Samuel P. (1968) 2006. *Political order in changing societies.* Renewed. Foreword by Francis Fukuyama. The Henry L. Stimson Lectures Series. Yale University Press.

———. 1993. "The clash of civilizations?" *Foreign Affairs* 72 (3): 22–49.

———. (1996) 2011. *The clash of civilisations and the remaking of world order.* Foreword by Zbigniew Brzezinski. Simon & Schuster.

Hurd, Ian. 2005. "Legitimacy." In *Encyclopedia of international relations and global politics,* edited by Griffiths Martin, 501–502. Routledge.

Hurrell, Andrew. 1990. "Kant and the Kantian Paradigm in International Relations." *Review of International Studies* 16 (3): 183–205.

Ah-hyun, Koo. 2021. *The role and limitations of future AI judges in the context of the impeachment of judges.* Local News. AiTIMES, February 1, 2021. Accessed June 10, 2023. https://www.aitimes.com/news/articleView.html?idxno=136139.

Iatrou, Evan. 2022. "A normative model of explanation for binary classification legal AI and its implementation on causal explanations of Answer Set Programming." In *Proceedings of the International Conference on Logic Programming 2022 (ICLP 2022),* edited by Joaquín Arias, Roberta Calegari, Luke Dickens, Wolfgang Faber, Jorge Fandinno, Gopal Gupta, Markus Hecher, et al. Haifa, Israel, July.

Ichikawa, Jonathan Jenkins, and Matthias Steup. 2018. "The analysis of knowledge." In *The Stanford Encyclopedia of Philosophy,* Summer 2018, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Iftimiei, Andra, and Mihai Iftimiei. 2022. "Law and IT technologies: Predictive justice." *Perspectives of Law and Public Administration* 11 (1): 169–175.

Ilegieuno, Sadiku, Okabonye Chukwuani, and Ifeoluwa Adaralegbe. 2022. "Artificial intelligence and the future of law practice in Nigeria." Chap. 14 in *Blockchain, artificial intelligence, and the internet of things: Possibilities and opportunities,* edited by Pethuru Raj, Ashutosh Kumar Dubey, Abhishek Kumar, and Pramod Singh Rathore, 307–325. EAI/Springer Innovations in Communication and Computing. Springer Cham. https://doi.org/10.1007/978-3-030-77637-4.

Izhikevich, Eugene M. 2003. "Simple model of spiking neurons." *IEEE Transactions on Neural Networks* 14 (6): 1569–1572.

Jacobsen, Jeppe T. 2022. "Cyberterrorism: Four reasons for its absence—so far." *Perspectives on Terrorism* 16 (5): 62–72.

Janiak, Andrew. 2021. "Newton's philosophy." In *The Stanford Encyclopedia of Philosophy,* Fall 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

John, Finnis. 2011. *Natural law and natural rights.* 2nd ed. Clarendon Law Series. Oxford University Press.

Jørgensen, Jørgen. 1937/1938. "Imperatives and logic." *Erkenntnis* 7:288–296.

Joutsen, Matti. 2010. "Legal traditions." Chap. 9 in *International crime and justice,* edited by Mangai Natarajan, Part II: Law, punishment, and crime control philosophies of the world, 67–74. Cambridge University Press. https://doi.org/10.1017/CBO9780511762116.014.

Kaltsas, Spyridon, Gerasimos Karoulas, Yiannis Karayiannis, and Fani Kountouri. 2022. "The political discourse of the church of Greece during the crisis: An empirical approach." *Religions* 13 (4). https://doi.org/10.3390/rel13040273.

Kamath, Uday, John Liu, and James Whitaker. 2019. *Deep learning for NLP and speech recognition.* Springer.

Kant, Immanuel. (1784) 1963. "Idea for a universal history from a cosmopolitan point of view." In *On history,* edited by Lewis White Beck, translated by Lewis White Beck, Rober E. Anchor, and Emil F. Fackenheim, 11–26. The Library of Liberal Art Series. Bobbs-Merrill.

———. (1795) 2006. "Toward perpetual peace: A philosophical sketch." In *Toward perpetual peace and other writings on politics, peace, and history,* edited by Pauline Kleingeld, translated by David L. Colclasure, 67–109. Rethinking the Western Tradition. Yale University Press.

Kausch, Kristina. 2022. "AI regulation in MENA: Brussels effect vs. Beijing effect." In *Liberty doom? Artificial intelligence in Middle Eastern security,* edited by Justine Leïla Belaïd. 27. Published by the European Institute of the Mediterranean, May.

Keeny, Ralph L. 1986. *Value-driven expert systems for decision support,* edited by Jeryl L. Mumpower, Ortwin Renn, Lawrence D. Phillips, and V. R. R. Uppuluri, Proceedings of the NATO Advanced Research Workshop on Expert Judgment and Expert Systems held in Porto, Portugal, August 25-29, 1986, 155–172. NATO Advanced Science Institutes (ASI) Series, Series F: Computer and systems sciences, vol.35. Springer-Verlag.

Kelsen, Hans. 1991. *General theory of norms.* Translated by Michael Hartney. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198252177.001.0001.

Kikarea, Eirini, and Maayan Menashe. 2019. "The global governance of cyberspace: reimagining private actors' accountability: introduction." *Cambridge International Law Journal* 8 (2): 153–170. https://doi.org/10.4337/cilj.2019.02.00.

Killeen, Molly. 2023. "German Constitutional Court strikes down predictive algorithms for policing" (February 16, 2023). Accessed March 13, 2023. https://www.euractiv.com/section/artificial-intelligence/news/german-constitutional-court-strikes-down-predictive-algorithms-for-policing/.

Kim, Pauline T. 2022. "Race-aware algorithms: Fairness, nondiscrimination and affirmative action." *California Law Review* 110 (5): 1539–1596. https://doi.org/10.15779/Z387P8TF1W.

Kim, Sung Ho. 2022. "Max Weber." In *The Stanford Encyclopedia of Philosophy,* Winter 2022, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Kissinger, Henry A. 2014. *World order.* Penguin Press.

———. 2018. *How the Enlightenment ends: Philosophically, intellectually — in every way — human society is unprepared for the rise of artificial intelligence.* Technology. The Atlantic, June. Accessed March 1, 2023. https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

Kissinger, Henry A., Eric Schmidt, and Daniel Huttenlocher. 2021. *The age of AI: And our human future.* Little, Brown and Company.

Kleingeld, Pauline. 1998. "Kant's cosmopolitan law: World citizenship for a global order." *Kantian Review* 2:72–90. https://doi.org/10.1017/S1369415400000200.

Knill, Christoph, and Jale Tosun. 2012. *Public policy: A new introduction.* Palgrave Macmillan.

Koch, Christof. 2016. *How the computer beat the Go master: As a leading go player falls to a machine, artificial intelligence takes a decisive step on the road to overtaking the natural variety.* Computing. Scientific American (SA), March 19, 2016. Accessed March 20, 2023. https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/.

Kodde, Claudia. 2016. "Germany's 'Right to be forgotten' – between the freedom of expression and the right to informational self-determination." *International Review of Law, Computers and Technology* 30 (1-2): 17–31. https://doi.org/10.1080/13600869.2015.1125154.

Kosa, David. 2015. "Selecting Strasbourg Judges: A Critique." Chap. 6 in *Selecting Europe's judges: A critical review of the appointment procedures to the European courts*, edited by Michal Bobek, 120–161. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198727781.003.0007.

Koshul, Basit Bilal. 2005. "Beyond the Enlightenment: Weber on the irreducible relationship between faith and science." Chap. 2 in *The postmodern significance of Max Weber's legacy: Disenchanting disenchantment*, 41–54. Palgrave Macmillan US. https://doi.org/10.1057/9781403978875_3.

Kuljanin, Dragan. 2019. "Why not a philosopher king? And other objections to Epistocracy." *Phenomenology and Mind* 16:80–89. https://doi.org/10.13128/Phe_Mi-26075.

Kumar, Sanjeev. 2020. "Impact of intellectuals and philosophers in French revolution 1789." *International Journal of History* 2 (1): 56–59.

Kusner, Matt J., Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. "Counterfactual fairness." In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 4069–4079. Long Beach, California, USA.

Labs of Latvia. 2021. *Artificial intelligence now in the courtroom*. May 17, 2021. Accessed March 3, 2023. https://labsoflatvia.com/en/news/artificial-intelligence-now-in-the-courtroom.

Lagioia, Francesca, Riccardo Rovatti, and Giovanni Sartor. 2023. "Algorithmic fairness through group parities? The case of COMPAS-SAPMOC." *AI and Society* 38 (2): 459–478.

Lamond, Grant. 2016. "Precedent and analogy in legal reasoning." In *The Stanford Encyclopedia of Philosophy*, Spring 2016, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Land, Molly K. 2020. "The problem of platform law: Pluralistic legal ordering on social media." Chap. 36 in *The Oxford handbook of global legal pluralism*, edited by Paul Schiff Berman, Part IX: Global legal pluralism and the deterritorialization of data, 975–994. Oxford University Press.

Lavrysen, Laurens. 2016. *Human rights in a positive state: Rethinking the relationship between positive and negative obligations under the European Convention on Human Rights*. Intersentia. https://doi.org/10.1017/9781780685311.

———. 2018. "Causation and positive obligations under the European Convention on Human Rights: A reply to Vladislava Stoyanova." *Human Rights Law Review* 18 (4): 705–718. https://doi.org/10.1093/hrlr/ngy027.

Law, Jonathan, ed. 2022. *A dictionary of law*. 10th ed. Oxford quick reference. Oxford University Press.

Layton, Roslyn. 2019. *The 10 Problems of the GDPR: The US can learn from the EU's mistakes and leapfrog its policy*. Statement before the Senate Judiciary Committee on the General Data Protection Regulation, March 12, 2019. https://www.judiciary.senate.gov/imo/media/doc/Layton%5C%20Testimony1.pdf.

Lazarus, Jeffrey V., Scott Ratzan, Adam Palayew, Francesco C. Billari, Agnes Binagwaho, Spencer Kimball, Heidi J. Larson, et al. 2020. "COVID-SCORE: A global survey to assess public perceptions of government responses to COVID-19 (COVID-SCORE-10)." *PLOS ONE* 15, no. 10 (October): 1–18. https://doi.org/10.1371/journal.pone.0240011. https://doi.org/10.1371/journal.pone.0240011.

Lefkowitz, David. 2020. *Philosophy and international law: A critical introduction*. Edited by Brian H. Bix and William A. Edmundson. Cambridge Introductions to Philosophy of Law. Cambridge University Press.

Leibowitz, Uri D., and Neil Sinclair. 2016. *Explanation in ethics and mathematics: Debunking and dispensability*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198778592.001.0001.

Leiter, Brian, and Michael Sevel. 2022. *Philosophy of law*. Online ed. Revised and updated by Jeannette L. Nolen. Encyclopedia Britannica, August 9, 2022. Accessed April 1, 2023. https://www.britannica.com/topic/philosophy-of-law.

Lemke, Coralie. 2020. *Une IA de Google surpasse les radiologues pour détecter le cancer du sein*. Sciences et avenir, January 3, 2020. Accessed January 10, 2023. https://www.sciencesetavenir.fr/sante/une-ia-de-google-surpasse-les-radiologues-pour-detecter-le-cancer-du-sein_140225.

Lemmens, Koen. 2015. "(S)electing judges for Strasbourg: A (dis)appointing process?" Chap. 5 in *Selecting Europe's judges: A critical review of the appointment procedures to the European Courts*, edited by Michal

Bobek, 95–119. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198727781.003.0006.

Letsas, George. 2006. "Two concepts of the margin of appreciation." *Oxford Journal of Legal Studies* 26 (4): 705–732. https://doi.org/10.1093/ojls/gql030.

Letwin, Jeremy. 2021. "Why completeness and coherence matter for the European Court of Human Rights." *European Convention on Human Rights Law Review* 2 (1): 119–154. https://doi.org/10.1163/26663236-bja10002.

Liese, Andrea, Jana Herold, Hauke Feil, and Per-Olof Busch. 2021. "The heart of bureaucratic power: Explaining international bureaucracies' expert authority." *Review of International Studies* 47 (3): 353–376. https://doi.org/10.1017/S026021052100005X.

Ma, Winnie, and Vincent Valton. 2023. *Toward an ethics of AI belief.* arXiv: 2304.14577 [cs.CY].

MacBride, Fraser. 2022. "Truthmakers." In *The Stanford Encyclopedia of Philosophy*, Fall 2022, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

MacCormick, Neil. 1992. "Legal deduction, legal predicates and expert systems." *International Journal for the Semiotics of Law* 5 (2): 181–202. https://doi.org/10.1007/BF01101868.

MacIntyre, Alasdair. (1981) 2007. *After virtue: A study in moral theory.* 3rd ed. Notre Dame Press.

Mantouvalou, Virginia. 2013. "Labour rights in the European Convention on Human Rights: An intellectual justification for an integrated approach to interpretation." *Human Rights Law Review* 13 (3): 529–555. https://doi.org/10.1093/hrlr/ngt001.

Markie, Peter, and M. Folescu. 2023. "Rationalism vs. empiricism." In *The Stanford Encyclopedia of Philosophy*, Spring 2023, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

McCarty, Thorne. 1977. "Reflection on TAXMAN: An experiment on artificial intelligence and legal reasoning." *Harvard Law Review* 90:837–893.

———. 1980. "The TAXMAN project: Towards a cognitive theory of legal argument." In *Computer science and law: An advanced course*, edited by Bryan Niblett, 23–43. Cambridge University Press.

McCulloch, Warren S., and Walter Pitts. 1943. "A logical calculus of the ideas immanent in nervous activity." *Bulletin of Mathematical Biophysics*, no. 5, 115–137.

McGrath, Thomas, Andrei Kapishnikov, Nenad Tomaev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. 2022. "Acquisition of chess knowledge in AlphaZero." *Proceedings of the National Academy of Sciences* 119 (47): e2206625119. https://doi.org/10.1073/pnas.2206625119.

Mchangama, Jacob, and Natalie Alkiviadou. 2021. "Hate speech and the European Court of Human Rights: Whatever happened to the right to offend, shock or disturb?" *Human rights law review* 21 (4): 1008–1042. https://doi.org/10.1093/hrlr/ngab015.

McKinney, Scott M., Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, et al. 2020. "International evaluation of an AI system for breast cancer screening." *Nature* (577): 89–94. https://doi.org/10.1038/s41586-019-1799-6.

McKinsey & Company. 2022. "What are Industry 4.0, the Fourth Industrial Revolution, and 4IR?" (April 17, 2022). Accessed May 2, 2023. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-are-industry-4-0-the-fourth-industrial-revolution-and-4ir#/.

Meester, Ronald, and Klaas Slooten. 2021. *Probability and forensic evidence: Theory, philosophy, and applications.* Cambridge University Press. https://doi.org/10.1017/9781108596176.

Miglietta, Marco, Nicolò Damiani, Gabriele Guerrini, and Francesco Graziotti. 2021. "Full-scale shake-table tests on two unreinforced masonry cavity-wall buildings: effect of an innovative timber retrofit." *Bulletin of Earthquake Engineering* 19 (6): 2561–2596. https://doi.org/10.1007/s10518-021-01057-5.

Mill, John Stuart. (1859) 1901. *On liberty.* The Walter Scott Publishing Co. https://www.gutenberg.org/ebooks/34901.

Minsky, Marvin, and Seymour A. Papert. 1969. *Perceptrons.* The MIT Press. https://doi.org/10.7551/mitpress/11301.001.0001.

Mokyr, Joel. 2004. *The gifts of Athena: Historical origins of the knowledge economy.* Princeton University Press.

Montesquieu. (1748) 1892. *De l'esprit des loix.* 2nd ed. Edited by Paul Janet. Books I-V. Libraire Ch. Delagrave.

Moore, Michael S. 2009. *Causation and responsibility: An essay in law, morals, and metaphysics.* Oxford university Press. https://doi.org/10.1093/acprof:oso/9780199256860.001.0001.

———. 2019. "Causation in the law." In *The Stanford Encyclopedia of Philosophy*, Winter 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Mori, Yuichi, Masashi Misawa, Jorge Bernal, Michael Bretthauer, Shin-ei Kudo, Amit Rastogi, and Gloria Fernández-Esparrach. 2022. "Artificial intelligence for disease diagnosis: the criterion standard challenge." *Gastrointestinal Endoscopy* 96 (2): 370–372. https://doi.org/10.1016/j.gie.2022.04.057.

Morris, William Edward, and Charlotte R. Brown. 2022. "David Hume." In *The Stanford Encyclopedia of Philosophy*, Summer 2022, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Mourtzis, Dimitris, John Angelopoulos, and Nikos Panopoulos. 2022. "A literature review of the challenges and opportunities of the transition from Industry 4.0 to Society 5.0." *Energies* 15 (17): 6276. https://doi.org/10.3390/en15176276.

Moyer, Christopher. 2016. *How Google's AlphaGo beat a Go world champion: Inside a man-versus-machine showdown.* Technology. March 28, 2016. Accessed May 1, 2023. https://www.theatlantic.com/technology/archive/2016/03/the-invisible-opponent/475611/.

MSI-AUT (CoE's committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence), Rapporteur: Karen Yeung. 2019. *Responsibility and AI: A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework.* Council of Europe study. DGI(2019)05. Printed at the Council of Europe.

Mulder, Dwayne H. n.d. *Objectivity.* Internet Encyclopedia of Philosophy. Accessed April 14, 2023. https://iep.utm.edu/objectiv/.

Mumpower, Jeryl L., Ortwin Renn, Lawrence D. Phillips, and V. R. R. Uppuluri, eds. 1986. "Expert judgment and expert systems," Proceedings of the NATO Advanced Research Workshop on Expert Judgment and Expert Systems held in Porto, Portugal, August 25-29, 1986. NATO Advanced Science Institutes (ASI) Series, Series F: Computer and systems sciences, vol.35. Springer-Verlag.

Munro, André. 2021. *State of nature.* Online ed. Political Theory. Encyclopedia Britannica, July 23, 2021. Accessed June 6, 2023. https://www.britannica.com/topic/state-of-nature-political-theory.

Nadler, Steven. 2022. "Baruch Spinoza." In *The Stanford Encyclopedia of Philosophy*, Summer 2022, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Navarro, Pablo E., and Jorge L. Rodríguez. 2014. *Deontic logic and legal systems.* Cambridge Introductions to Philosophy and Law. Cambridge University Press. https://doi.org/10.1017/CBO9781139032711.

Nickel, James. 2021. "Human Rights." In *The Stanford Encyclopedia of Philosophy*, Fall 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Niler, Eric. 2019. *Can AI be a fair judge in court? Estonia thinks so: Estonia plans to use an artificial intelligence program to decide some small-claims cases, part of a push to make government services smarter.* Wired, March 25, 2019. Accessed February 20, 2023. https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/.

Nishigai, Konatsu. 2021. "Two types of formalism of the rule of law." *Oxford Journal of Legal Studies* 42 (2): 495–520. https://doi.org/10.1093/ojls/gqab039.

Nitta, Katsumi, and Ken Satoh. 2020. "AI applications to the law domain in Japan." *Asian Journal of Law and Society* 7 (3): 471–494. https://doi.org/10.1017/als.2020.35.

Nolan, Donal. 2013. "Negligence and human rights law: The case for separate development." *The Modern Law Review* 76 (2): 286–318. https://doi.org/10.1111/1468-2230.12013.

Nussberger, Angelika. 2020. *The European Court of Human Rights.* 1st ed. (online). Edited by Mark Janis, Douglas Guilfoyle, Stephan Schill, Bruno Simma, and Kimberley Trapp. Elements of International Law. Oxford University Press. https://doi.org/10.1093/law/9780198849643.001.0001.

OpenAI, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, et al. 2019. *Dota 2 with large scale deep reinforcement learning.* arXiv: 1912.06680 [cs.LG].

Paine, Thomas. (1776) 2021. *Common sense: Addressed to the inhabitants of America.* W. & T. Bradford, August 10, 2021. https://www.gutenberg.org/files/147/147-h/147-h.htm.

Palm, Rasmus Berg, Ulrich Paquet, and Ole Winther. 2018. "Recurrent relational networks." In *Proceedings of the 32nd International Conference on Neural Information Processing Systems,* 3372–3382. NIPS'18. Montréal, Canada: Curran Associates Inc.

Park, H. A. 2016. "Are we ready for the fourth industrial revolution?" *Yearbook of medical informatics* 1:1–3. https://doi.org/10.15265/IY-2016-052.

Peter, Fabienne. 2017. "Political legitimacy." In *The Stanford Encyclopedia of Philosophy,* Summer 2017, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Philbeck, Thomas, and Nicholas Davis. 2018. "The fourth industrial revolution: Shaping a new era." *Journal of International Affairs* 72 (1): 17+.

Pilotti, Maura A. E., Halah Al Kuhayli, and Eman Abdulhadi. 2021. "Judging the misdeeds of others: A study of embodied cognition in the Middle East." *The International Journal of Diverse Identities* 21 (1): 1–12. https://doi.org/10.18848/2327-7866/CGP/v21i01/1-12.

Plattner, Marc F. 2019. "Illiberal democracy and the struggle on the right." *Journal of Democracy* 30 (1): 5–19. https://doi.org/10.1353/jod.2019.0000.

Poggi, Francesca. 2021. "Defeasibility, law, and argumentation: A critical view from an interpretative standpoint." *Argumentation* 35 (3): 409–434. https://doi.org/10.1007/s10503-020-09544-w.

Popper, Karl. 2005. *The logic of scientific discovery.* Routledge.

———. 2012. *The open society and its enemies.* Routledge.

Pranab, Mukhopadhyay, Nawn Nandan, and Das Kalyan. 2017. *Global change, ecosystems, sustainability: Theory, methods, practice.* Sage Publications Pvt. Ltd.

Press Service of the ECtHR. 2018. *Factsheet on extra-territorial jurisdiction of States Parties.* July. Accessed May 13, 2023. https://www.echr.coe.int/Pages/home.aspx?p=press/factsheets&c.

———. 2022a. *Factsheet on mass surveillance.* September. Accessed April 4, 2023. https://www.echr.coe.int/Pages/home.aspx?p=press/factsheets&c.

———. 2022b. *Factsheet on surveillance at workplace.* December. Accessed April 4, 2023. https://www.echr.coe.int/Pages/home.aspx?p=press/factsheets&c.

Putnam, Hilary. 2002. *The collapse of the fact/value dichotomy and other essays.* Harvard University Press.

Rai, Arun. 2020. "Explainable AI: from black box to glass box." *Journal of the Academy of Marketing Science* 48 (1): 137–141. https://doi.org/10.1007/s11747-019-00710-5.

Ralston, Shane J. n.d. *American Enlightenment Thought.* Internet Encyclopedia of Philosophy. Accessed April 14, 2023. https://iep.utm.edu/american-enlightenment-thought/.

Ratner, Claudia. 2021. "When "Sweetie" is not so sweet: Artificial intelligence and its implications for child pornography." *Family Court Review* 59 (2): 386–401. https://doi.org/10.1111/fcre.12576.

Raymond, Mark. 2013. "Puncturing the myth of the internet as a commons." *Georgetown Journal of International Affairs. International engagement on Cyber III: State building on a new frontier (2013-14),* 53–64.

Raz, Joseph. 1979. *The authority of law: Essays on law and morality.* Oxford University Press.

Read, Thornton. 1961. *Command and control.* Policy Memorandum, No. 24. Center of International Studies. Woodrow Wilson School of Public and International Affairs. Princeton University, June 15, 1961.

Remus, Dana, and Frank S. Levy. 2016. *Can robots be lawyers? Computers, lawyers, and the practice of law.* Available at SSRN, November 27, 2016. http://dx.doi.org/10.2139/ssrn.2701092.

Richardson Oakes, Anne, and Haydn Davies. 2016. "Justice must be seen to be done: A contextual reappraisal." *Adelaide Law Review* 37 (2): 461–494.

Roberts, Huw, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. 2021. "The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation." *AI & Society* 36 (1): 59–77. https://doi.org/10.1007/s00146-020-00992-2.

Robertson, David. 2004. *A dictionary of human rights.* 2nd ed. Edited by Paul Kelly. Europa Publications.

Rogerson, Annys, Emma Hankins, Pablo Fuentes Nettel, and Sulamaan Rahim. 2022. *Government AI readiness index 2022.* Edited by Kirsty Trim and Sulamaan Rahim. Oxford Insights.

Rousseau, Jean-Jacques. (1762) 1920. *The social contract and discourses.* Edited by Ernest Rhys. J. M. Dent & Sons.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning representations by back-propagating errors." *Nature* 323 (6088): 533–536. https://doi.org/10.1038/323533a0.

Russell, Stuart J., Peter Norvig, Ming-Wei Chang, Jacob Devlin, Anca Dragan, David Forsyth, Ian Goodfellow, et al. 2021. *Artificial Intelligence: A modern approach.* 4th ed. Gobal ed. Edited by Stuart Russell and Peter Norvig. Pearson series in artificial intelligence. Pearson.

Russo, Federica. 2022. *Techno-scientific practices: An informational approach.* Rowman & Littlefield Publishers.

Sample, Ian. 2020. *AI system outperforms experts in spotting breast cancer.* The Guardian, January 1, 2020. Accessed January 10, 2023. https://www.theguardian.com/society/2020/jan/01/ai-system-outperforms-experts-in-spotting-breast-cancer.

Sartor, Giovanni, and Andrea Loreggia. 2020. *The impact of algorithms for online content filtering or moderation - upload filters.* Study requested by the JURI Committee. European Parliament Think Tank. Accessed November 1, 2022. https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2020)657101.

Scheeck, Laurent. 2005. "Solving Europe's binary human rights puzzle: The interaction between supranational courts as a parameter of European governance." *Questions de recherche/Research Questions, Centre d'études et de recherches internationales (CERI-Sciences Po/CNRS),* no. 15 (October).

Schliesser, Eric, and Tamás Demeter. 2020. "Hume's Newtonianism and Anti-Newtonianism." In *The Stanford Encyclopedia of Philosophy,* Summer 2020, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Schlosser, Markus. 2019. "Agency." In *The Stanford Encyclopedia of Philosophy,* Winter 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Schroeder, Mark. 2021. "Value theory." In *The Stanford Encyclopedia of Philosophy,* Fall 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Schroeter, Francois, Laura Schroeter, and Kevin Toh. 2020. "A new interpretivist metasemantics for fundamental legal disagreements." *Legal Theory* 26 (1): 62–99. https://doi.org/10.1017/S1352325220000063.

Schwab, Klaus. 2016. *The fourth industrial revolution: What it means and how to respond.* Foreign Affairs. Accessed May 20, 2023. https://www.foreignaffairs.com/world/fourth-industrial-revolution.

———. 2023. *The fourth industrial revolution.* Encyclopedia Britannica, March 24, 2023. Accessed May 1, 2023. https://www.britannica.com/topic/The-Fourth-Industrial-Revolution-2119734.

Schwitzgebel, Eric. 2021. "Belief." In *The Stanford Encyclopedia of Philosophy,* Winter 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Searle, John. 1980. "Minds, brains, and programs." *Behavioral and Brain Sciences* 3:417–424.

Sedley, Stephen. 2015. *Lions under the throne: Essays on the history of English public law.* Cambridge University Press.

Sen, Amartya K. 1967. "The nature and classes of prescriptive judgments." *The Philosophical Quarterly* 17 (66): 46–62. https://doi.org/10.2307/2218365.

Serebrin, Jacob. 2023. *Quebec man who created synthetic, AI-generated child pornography sentenced to prison.* Montreal: Canadian Broadcasting Corporation (CBC), April 26, 2023. Accessed May 19, 2023. https://www.cbc.ca/news/canada/montreal/ai-child-abuse-images-1.6823808.

Shafer-Landau, Russ. 2010. "The possibility metaethics." *Boston University Law Review* 90:479–496.

Shaftesbury, Lord. (1711) 2000. *Shaftesbury: Characteristics of Men, Manners, Opinions, Times.* Edited by Lawrence E. Klein. Cambridge Texts in the History of Philosophy. Cambridge University Press. https://doi.org/10.1017/CBO9780511803284.

Shapiro, Stewart. 2009. "We hold these truths to be self-evident: But what do we mean by that?" *The review of symbolic logic* 2 (1).

Shecaira, Fábio P. 2011. "Hume and noncognitivism." *History of Philosophy Quarterly* 28 (3): 267–287.

Shue, Henry. 1980. *Basic rights: Subsistence, affluence, and U.S. foreign policy.* Princeton University Press.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529 (7587): 484–489. https://doi.org/10.1038/nature16961.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al. 2018. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." *Science* 362 (6419): 1140–1144. https://doi.org/10.1126/science.aar6404.

Singh, Sudhvir, Christine McNab, Rose McKeon Olson, Nellie Bristol, Cody Nolan, Elin Bergstrøm, Michael Bartos, et al. 2021. "How an outbreak became a pandemic: a chronological analysis of crucial junctures and international obligations in the early months of the COVID-19 pandemic." *The Lancet* 398 (10316): 2109–2124. https://doi.org/10.1016/S0140-6736(21)01897-3.

Smith, George. 2008. "Newton's Philosophiae Naturalis Principia Mathematica." In *The Stanford Encyclopedia of Philosophy*, Winter 2008, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Sophocles. (5th century BC) 1984. *The three Theban plays: Antigone, Oedipus the king, Oedipus at Colonus.* Reprint. Translated by Robert Fagles. Introduction and notes by Bernard Knox. Penguin Books.

Spaak, Torben. 2009. "Explicating the concept of legal competence." In *Concepts in law*, edited by Jaap C. Hage and Dietmar von der Pfordten, 88:67–80. Law and Philosophy Library. Springer.

Spinoza, Benedict. (1677) 2017. *The Ethics (orig: Ethica ordine geometrico demonstrata).* Translated by R. H. M. Elwes. Project Gutenberg, December 11, 2017. https://www.gutenberg.org/files/3800/3800-h/3800-h.htm.

Stacey, Stephanie. 2023. *'Robot lawyer' DoNotPay is being sued by a law firm because it 'does not have a law degree'.* Insider, March 12, 2023. Accessed March 13, 2023. https://www.businessinsider.com/robot-lawyer-ai-donotpay-sued-practicing-law-without-a-license-2023-3?international=true&r=US&IR=T.

Stanton-Ife, John. 2022. "The limits of law." In *The Stanford Encyclopedia of Philosophy*, Spring 2022, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Stavropoulos, Nicos. 1996. *Objectivity in law.* Oxford University Press.

Steffen, Will, Angelina Sanderson, Peter Tyson, Jill Jäger, Pamela Matson, Berrien Moore, Frank Oldfield, et al. 2005. *Global change and the earth system: A Planet Under Pressure.* Global Change - The IGBP Series. Springer. https://doi.org/10.1007/b137870.

Steinberg, Justin. 2022. "Spinoza's Political Philosophy." In *The Stanford Encyclopedia of Philosophy*, Summer 2022, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Stoyanova, Vladislava. 2018. "Causation between state omission and harm within the framework of positive obligations under the European Convention on Human Rights." *Human Rights Law Review* 18 (2): 309–346. https://doi.org/10.1093/hrlr/ngy004.

Sun, Ron. 2014. "Connectionism and neural networks." Chap. 5 in *The Cambridge handbook of artificial intelligence*, edited by Keith Frankish and William M. Ramsey, Part II: Architecture. Cambridge University Press.

Sunder, Madhavi. 2020. "Fighting fundamentalism with pluralism: Technologies of Enlightenment during the Arab Spring." Chap. 37 in *The Oxford handbook of global legal pluralism*, edited by Paul Schiff Berman, Part IX: Global legal pluralism and the deterritorialization of data, 995–1015. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197516744.013.13.

Tasioulas, John. 2011. "On the nature of human rights." In *The philosophy of human rights: Contemporary controversies*, edited by Gerhard Ernst and Jan-Christoph Heilinger, Part II: Human rights: political or moral, 17–60. De Gruyter.

Tate, C. Neal. 2023. *Judicial review.* Encyclopedia Britannica, January 5, 2023. Accessed June 2, 2023. https://www.britannica.com/topic/judicial-review.

Taylor, Katherine Fischer. 2013. "Geometries of power: Royal, revolutionary, and postrevolutionary French courtrooms." *Journal of the Society of Architectural Historians* 72 (4): 434–474.

Taylor, Luke. 2023. "Colombian judge says he used ChatGPT in ruling: Juan Manuel Padilla asked the AI tool how laws applied in case of autistic boy's medical funding, while also using precedent to support his decision." (Bogotá) (February 3, 2023). Accessed March 11, 2023. https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling.

Thoma, Johanna. 2022. "Risk Imposition by artificial agents: The moral proxy problem." Chap. 4 in *The Cambridge handbook of artificial intelligence: Interdisciplinary perspectives*, edited by Silja Voeneky, Philipp Kellmeyer, Oliver Mueller, and Wolfram Burgard, Part II: Foundations of responsible AI. Cambridge University Press.

Toosi, Amirhosein, Andrea Bottino, Babak Saboury, Eliot Siegel, and Arman Rahmim. 2021. "A brief history of AI: How to prevent another winter (A critical review)." *PET Clinics* 16 (4): 449–469. https://doi.org/10.1016/j.cpet.2021.07.001.

Tuulik, Maria-Elisa (Public relations advisor/*ee*: Avalike suhete nõunik). 2022. *Estonia does not develop AI Judge.* Ministry of Justice, Republic of Estonia, February 16, 2022. Accessed February 20, 2023. https://www.just.ee/en/news/estonia-does-not-develop-ai-judge.

Ulenaers, Jasper. 2020. "The impact of artificial intelligence on the right to a fair trial: Towards a robot judge?" *Asian Journal of Law and Economics* 11 (2): 20200008. https://doi.org/doi:10.1515/ajle-2020-0008.

van Roojen, Mark. 2018. "Moral cognitivism vs. non-cognitivism." In *The Stanford Encyclopedia of Philosophy*, Fall 2018, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Vibert, Frank. 2007. *The rise of the unelected: Democracy and the new separation of powers.* Cambridge University Press.

Vidgen, Bertie, Emily Burden, and Helen Mergetts. 2021. *Understanding online hate: VSP (video-sharing platforms) regulation and the broader context.* The Alan Turing Institute.

von Wright, Georg Henrik. 1951. "Deontic logic." *Mind* 60:1–15.

Voronov, Maxim, and Klaus Weber. 2020. "People, actors, and the humanizing of institutional theory." *Journal of Management Studies* 57 (4): 873–884. https://doi.org/https://doi.org/10.1111/joms.12559.

Voskuhl, Adelheid. 2013. *Androids in the Enlightenment: Mechanics, artisans, and cultures of the self.* University of Chicago Press. https://doi.org/10.7208/chicago/9780226034331.001.0001.

Waldron, Jeremy. 2020. "The rule of law." In *The Stanford Encyclopedia of Philosophy*, Summer 2020, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Walsh, Alistair. 2017. *Saudi Arabia grants robot citizenship.* Deutsche Welle (DW), October 28, 2017. Accessed June 14, 2023. https://www.dw.com/en/saudi-arabia-grants-citizenship-to-robot-sophia/a-41150856.

Walton, Douglas. 2002. *Legal argumentation and evidence.* The Pennsylvania State University Press.

Wang, Quan-Jing, Gen-Fu Feng, Hai-Jie Wang, and Chun-Ping Chang. 2021. "The impacts of democracy on innovation: Revisited evidence." *Technovation* 108:102333. https://doi.org/https://doi.org/10.1016/j.technovation.2021.102333.

Wearden, Graeme. 2016. *Davos 2016: eight key themes for the World Economic Forum: Political and business leaders gather at Swiss ski resort to discuss issues including robots, terrorism, migration and inequality.* Davos: The Guardian, January 19, 2016. Accessed June 8, 2023. https://www.theguardian.com/business/2016/jan/19/world-economic-forum-davos-2016-eight-key-themes-robotics-migration-markets-climate-change-europe-medicine-inequality-cybercrime.

Weart, Spencer R. 2023. "The climate commons and the survival of democracy." Chap. 11 in *How democracy survives: Global challenges in the Anthropocene*, edited by Michael Holm and R. S. Deese, Part III: Confronting the Anthropocene, 179–195. Democratization and Autocratization Studies. Routledge.

Weber, Max. (1920) 1947. *The theory of social and economic organization.* Edited by Talcott Parsons. Translated by A. M. Henderson and Talcott Parsons. The Free Press.

———. 1974. *From Max Weber: Essays in sociology.* Reprint. Edited and translated by H. H. Gerth and G. Wright Mills. Routledge and Kegan Paul.

WEF (World Economic Forum). 2019. *Fourth industrial revolution: Beacons of technology and innovation in manufacturing.* Whitepaper. In collaboration with McKinsey & Company. January 10, 2019. Accessed May 23, 2023. https://www.weforum.org/whitepapers/fourth-industrial-revolution-beacons-of-technology-and-innovation-in-manufacturing/.

Weston, Burns H. 2023. *Human rights.* Encyclopedia Britannica, May 18, 2023. Accessed May 20, 2023. https://www.britannica.com/topic/human-rights.

Winter, Christoph, Nicholas Hollman, and David Manheim. 2023. "Value alignment for advanced artificial judicial intelligence." *American Philosophical Quarterly* 60 (2): 187–203. https://doi.org/10.5406/21521123.60.2.06.

Winter, Jack. 2016. "Justice for hedgehogs, conceptual authenticity for foxes: Ronald Dworkin on value conflicts." *Res Publica* 22 (4): 463–479. https://doi.org/10.1007/s11158-015-9285-y.

WJP (World Justice Project). 2022. *Rule of Law Index® 2022.* Published by the World Justice Project (WJP). https://worldjusticeproject.org/rule-of-law-index/global.

Yang, Zhun, Adam Ishay, and Joohyung Lee. 2020. "NeurASP: Embracing neural networks into Answer Set Programming." In *Proceedings of the twenty-ninth international joint conference on artificial intelligence*, edited by Christian Bessiere, 1755–1762. International joint conferences on artificial intelligence organization. https://doi.org/10.24963/ijcai.2020/243.

Zakaria, Fareed. 1997. "The rise of illiberal democracy." *Foreign Affairs* 76:22–43.

Zhabina, Alena. 2023. *How China's AI is automating the legal system.* Deutsche Welle (DW), January 20, 2023. Accessed March 2, 2023. https://www.dw.com/en/how-chinas-ai-is-automating-the-legal-system/a-64465988.

Zharova, Anna, Vladimir Elin, and Peter Panfilov. 2019. "Introducing artificial intelligence into law enforcement practice: The case of Russia." In *Proceedings of the 30th International DAAAM Symposium "Intelligent Manufacturing and Automation"*, edited by Branko Katalinic, 30:688–692. 1. DAAAM International. https://doi.org/10.2507/30th.daaam.proceedings.094.

CHAPTER II

Prickles of hedgehogs and skulks of foxes
Towards a new philosophy of science

ART BY NICO MAVRIDI

Philosophers, political and legal scientists, politicians, journalists and all the others that are *de facto* in the forefront of the discourse that shapes the content of legitimacy pillars like human rights, democracy, and the rule of law, they do not usually have the technical background required to comprehend, let alone to participate in the engineering of AlgoAI models.[1] At the same time, AI engineers may have mastered applied mathematics, computer and data science, python & r, but they usually lack the theoretical background of the aforementioned "*legitimacy*" experts. In his 2018 article, Kissinger diagnosed this chasm between the two groups of experts urging at the conclusion of his article for the formation of a committee of "*eminent thinkers*" that will prevent Enlightenment's foretold death, implying that those thinkers should be from both experts camps (*cf.* CEPEJ 2019, p.9).[2] The above observation brings forward the questions of *who* should be part of such an interdisciplinary team, *which* tasks each group of experts should resolve, *how* should those experts co-operate to fulfill those tasks, and *how* those tasks can/should be evaluated. The first question was answered in chapter i: Al-

[1]Take for instance the value of the *rule of law*. As Waldron argues, the rule of law is a "*working*" political concept being shaped by "*ordinary citizens, lawyers, activists and politicians as of the jurists and philosophers*" (Waldron 2020, §2). Clearly, none of them is required to have knowledge about AI to partake in the evolution of rule of law's content. *Cf.* with the group of individuals that conferred at the Congress of Europe (1948) to set the foundations for the CoE, the Convention, & the ECtHR: politicians, members of European parliaments and governments, representatives from employers' organisations and trade unions, journalists, and intellectuals (§introduction).

[2]Kissinger had firstly applied this advice to himself. The inspiration for his 2018 Atlantic article was a talk Kissinger accidentally attended in 2015 whose topic was about an AI model that plays the game Go (*cf.* §I.3.2.1.1, ¶1). Being self-"*[a]ware of [his] lack of technical competence*", Kissinger organised informal dialogues about AI in collaboration with experts from *both* "*technology and the humanities*" (*ibid.*). Kissinger ended up co-authoring a book ("*The age of AI: And our human future*", 2021) and another article published in the Atlantic ("*Metamorphosis*", 2019) with two such experts, Eric Schmidt (former CEO of Google (2001-2011) and executive chairman of both Google (2011-2015) and Alphabet Inc. (2015-2017)) and Daniel Huttenlocher (founder of Cornell Tech and current dean of the MIT Schwarzman College of Computing). The article was more or less a preamble of the book, while both of them were a continuation of Kissinger's 2018 article on Enlightenment's foretold death.

GOAI engineers should consist of *AI engineers*, legitimacy experts including *legal scientists*,[3] *logicians & formal philosophers* (§I.2.5, ¶9; §I.4). In this chapter, I answer the second and third questions, while in CHAPTER III, I will address the fourth one.

More precisely, in §1 of this chapter, I introduce Isaiah Berlin's (1953) famous allegorical dichotomy between *fox & hedgehog* thinkers. I will use this allegory to make sense of the collaborative practice among the legal ALGOAI engineers. In §2, I introduce a typology of *disciplinarities* (*inter*-disciplinarity, *cross*-disciplinarity, *multi*-disciplinarity, etc) arguing that legal ALGOAI engineering is, in reality, a *cross*-disciplinary practice and not an *inter*-disciplinary one. I further argue by using the fox-hedgehog analogy how the different disciplinary experts should co-operate in a *meta*-disciplinary level to engineer legitimate ALGOAI models. I continue in §3 by placing ALGOAI engineering in the context of *philosophy of interdisciplinarity*, the nascent evolutionary stage in philosophy of science. I use this contextualisation to explicate two core aspects of ALGOAI engineering: (α) its *trans*-disciplinarity. I.e., the collaboration among disciplines so as to produce knowledge & ontology that *transcends* academic disciplines to realise non-academically oriented ends (political, legal, social, etc); (β) the nuances & dangers that lurk during the production of *contactual information* by the collaboration of the different disciplines. To do so, I use a classical example in the discipline of AI & law, Danziger, Levav, and Avnaim-Pesso's landmark & controversial 2011 paper about the so-called *hungry judge* effect (*cf.* §I.3.1, ¶1). Finally, in §4, by using once more the fox-hedgehog schema, I argue how *logic* glues the three types of experts involved in ALGOAI engineering (AI engineers, legal scientists, and logicians & formal philosophers) in the *current* legal ALGOAI engineering practice, as well as how it *should* glue them in order to satisfy the legitimacy requirements set in CHAPTER I.

II.1 On foxes and hedgehogs

πολλ᾽ οἶδ᾽ ἀλώπηξ, ἀλλ᾽ ἐχῖνος ἕν, μέγα, Ἀρχίλοχος

Isaiah Berlin was a Latvian-born (1909–1997) Russian-Jewish intellectual powerhouse (Cherniss and Hardy 2022, §Introduction and §1). Despite being one of the founders of a novel discipline,[4] a pioneer of the so-called Oxford philosophy popularising it to the USA, and despite his distinguished work in a diverse range of disciplines, from philosophy and history to political science and Russian literature (*ibid.*, §1), for a large portion of the public he is known for an analogy that he made on the first *two* (!) pages of his 1953 "*The hedgehog and the fox: An essay on Tolstoy's view of history*". An analogy which was meant more as "*a kind of enjoyable intellectual game*" than a serious argument (Berlin and Jahanbegloo 1991, p.188). More precisely, Berlin borrows the Greek soldier-poet Archilochus'[5] line "*Πολλ᾽ οἶδ᾽ ἀλώπηξ, ἀλλ᾽ ἐχῖνος ἕν, μέγα.*" (*transl:* "*The fox knows many things, but the hedgehog knows one big thing.*") to divide thinkers into two groups: *hedgehogs* and *foxes*. Hedgehogs are those "*centripetal*" thinkers whose work is centered around one "*single, universal, organising principle*", what Archilochus calls "*one big thing*". Foxes are those "*centrifugal*" thinkers whose work consists of diverse and sometimes contradictory interests that are not bounded by any non-circumstantial organising principle. For Berlin, examples of hedgehogs are Dante, Plato, Lucretius, Pascal, Hegel, Dostoevsky, Nietzsche, Ibsen, Proust, while examples of foxes are Shakespeare, Aristotle, Herodotus, Montaigne, Erasmus, Molière, Goethe, Pushkin, Balzac, Joyce (Berlin (1953) 1954, pp.1-2). When Berlin commented 38 years on his famous dichotomy, he rightly clarified its fluidity; while there are dedicated foxes and hedgehogs, some thinkers are *neither* and some are *both* (Berlin and Jahanbegloo 1991, p.189).

Berlin's figurative analogy has been (ab)used by multiple disciplines since its publication. It has been used in business and economics, chemistry, political, medical, cognitive, and social sciences, law, history and (meta)-philosophy, strategy and international relations,[6] even in a Woody Allen film.[7] I will (ab)use it one more

[3] For what I construe as *legitimacy experts, see* §I.2.5, ¶9. Since in this Thesis I have focused on legal ALGOAI, for reasons of convenience, from all non-logicians/formal philosopher legitimacy experts, I will be concerned only with *legal scientists*. To the infuriated philosopher of science, I kindly ask to be patient until §4.1.2 where I define what is "*legal science*".

[4] The discipline of *intellectual history* (Britannica 2023), which is a "*branch of history that deals with the historical propagation and dissemination of ideas* " (Vann 2023).

[5] Johnson 1996, p.326. [Archilochus (Ἀρχίλοχος) c. 680 – c. 645 BC, born in the Greek island of Paros]

[6] Singh 2010; Crow 2020; Mak 2012; Gould 2003; Kampen 2020; Harpham 2015; Winter 2016.

[7] In the 1992 film "*Husbands and Wives*", the wife, Sally, confesses to her therapists that during a sexual intercourse with her extramarital lover, instead of enjoying the moment, she was enumerating in her thoughts the people in her life who are *foxes* and those who are *hedgehogs*. For Johnson 1996, this is a way for Sally to rationalise and distance herself from the sexual experience as a defense reflex induced by past sexual abuse.

time. Before doing so, shout-out to Ronald Dworkin's[8] 2011 swansong "*Justice for hedgehogs*" which is on the *epicenter* of the objectivity challenge. *Contra* the dominant fox-approach which argues that *values* are in principle heterogeneous and conflict with each other in many real-life situations (e.g., the value of human rights conflicting with the value of democracy (§I.2.8, ¶2)), for Dworkin there exists a *coherent* complementary unity among them (*see also* the *coherence adequacy requirement* in §III.3.2.1). Berlin's hedgehog works towards a "*single central vision*", a "*coherent*" system which is bound together by a "*single, universal, organising principle*" (Berlin 1954, p.1). For Dworkin, that "*coherent system*" is the foregoing coherent system of values, and that "*organising principle*" is his method of *interpreting concepts* that I am using adjusted in §III.3.3. A method that Dworkin first introduced in his groundbreaking "*Law's Empire*" (1986) and he finalised in his "*Justice for hedgehogs*".

Amidst this mosaic of uses, the hedgehog-fox analogy has also been used to make sense of *interdisciplinarity*.[9] Such is the use of the analogy that I will also make, talking about *foxes* as *gluons*. Before doing so, allow me to make an intermediate stop to differentiate between *interdisciplinarity* and other *X*-disciplinarities: *meta*-disciplinarity, *cross*-disciplinarity, and the like.

II.2 *Meta*-disciplinarity

II.2.1 *Inter*-disciplinarity, *cross*-disciplinarity, and the like

A classical problem in metaphysics is the so-called *one-and-the-many problem*. Assume an object (e.g., a chair) and its parts (e.g., legs, arms, back). The one-and-the-many problem poses the question of *how* does the multiplicity of the object's parts "*produce a unity*",[10] the unity being the object. "*How do the parts conspire to form a whole? What is the difference between a unity and a mere congeries?*" (Priest 2014, p.xvi). For some, the answer is that there must be *something* that binds these parts (the *many*) together so as to form the unity (the *one*), what metaphysician Graham Priest coined as *gluon* (*ibid.*, p.9). A similar challenge appears when one asks *what interdisciplinarity is*. An *intuitive* answer is that interdisciplinarity is the collaboration among experts from *multiple* disciplines so as to achieve a final *coherent* output (*cf.* Arnold 2020, p.1445). Once again, we are brought before the one-and-the-many question: what differentiates a mere congeries of disciplinary practices from a unified coherent interdisciplinary collaboration? Is there an equivalent of a gluon?

mere congeries *v.* *glued* congeries

In the 1972 paper "*Towards interdisciplinarity and transdisciplinarity in education and innovation*", a foundational paper for the contemporary conception of interdisciplinarity,[11] Jantsch answered the one-and-the-many question by providing demarcation criteria to discern different ways of *organising* disciplinary practices like the *inter*-disciplinary way, the *cross*-disciplinary, the *pluri*-disciplinary, and more. As we will see later on, legal ALGOAI engineering is after all *not* an *inter*-disciplinary practice, but a *cross*-disciplinary one. But first things first, let's see how Jantsch's typology of *X*-disciplinarities answers the one-and-the-many question. The first difference between a "*mere congeries of disciplinary practices*" and a "*unified coherent*" collaboration is that in the latter the disciplinary experts exercise their disciplinary research towards the realisation of *common* ends. Take for instance the classical disciplinary segmentation of universities in different departments (e.g., department of physics, psychology, law) originating from 19th century Berlin, Germany[12], a characteristic example of multiple disciplinary research environments (the departments) existing in the same research environment (the university), but operating in parallel isolation each pursuing *its own separate* disciplinary ends. This is the case of *multi*-disciplinary practice or simply *multi-disciplinarity* (henceforth MULTIDI). Sometimes, the experts of a specific discipline (e.g., criminal law) need to collaborate with the experts of other disciplines (e.g., medical examiners that perform autopsies, forensic scientists that perform DNA tests or track digital traces) so as to achieve a specific disciplinary end (e.g., successfully defend a client before a court). This is the case of a *cross*-disciplinary practice or simply *cross-disiplinarity* (henceforth CROSSDI). If experts of a discipline *X* work towards satisfying ends of another discipline *Y*, I will call the former discipline *dominated discipline* and the latter *dominating discipline*. Finally, some other times, disciplinary experts collaborate so as to achieve *new* ends that are *not* of

multi-, *cross*-, and *inter*-disciplinarities

[8] Ronald Dworkin (1931-2013; *see* Dworkin 1986, 2011 and Waluchow and Sciaraffa 2016 for an Oxford overview of his legacy) is one of the foundational figures of contemporary jurisprudence, next to philosophers like H. L. A. Hart (Hart 1961; Hart and Honoré 1985), John Rawls (Rawls 1999, 2000), and Joseph Raz (Raz 1979). The cited bibliography is some of their publications used for this Thesis.

[9] *See e.g.* Morgan's 2018 "*Forensic science needs both the 'hedgehog' and the 'fox'*", a paper about how interdisciplinary research in forensic science, a *paradigmatic* interdisciplinary discipline, should look like.

[10] Priest 2014, p.xv.

[11] Schmidt 2022, pp.1,16; Thorén and Persson 2013, p.1857. Jantsch's paper was popularised in the 1972 OECD Paris conference "*Interdisciplinarity: Problems of teaching and research in universities*" that is considered to be the foundational conference for the contemporary conception of the concept of interdisciplinarity (Schmidt 2022, p.16; *cf.* Arnold 2020, pp.1444-1445). Jantsch had previously published another version of the paper in 1970 that was reprinted in 1972 in the "*Higher Education*" journal. For different approaches on the origin of the term "*interdisciplinarity*" *see* Schmidt 2022, §1, endnote 3 and Arnold 2020, pp.1455-1456.

[12] Campbell and Schneider 2020, p.1446.

a specific discipline. Those ends are goals in a *meta*-level outside of the experts' disciplines. This is the case of *inter*-disciplinary practice or simply *inter-disciplinarity* (henceforth *ID*).

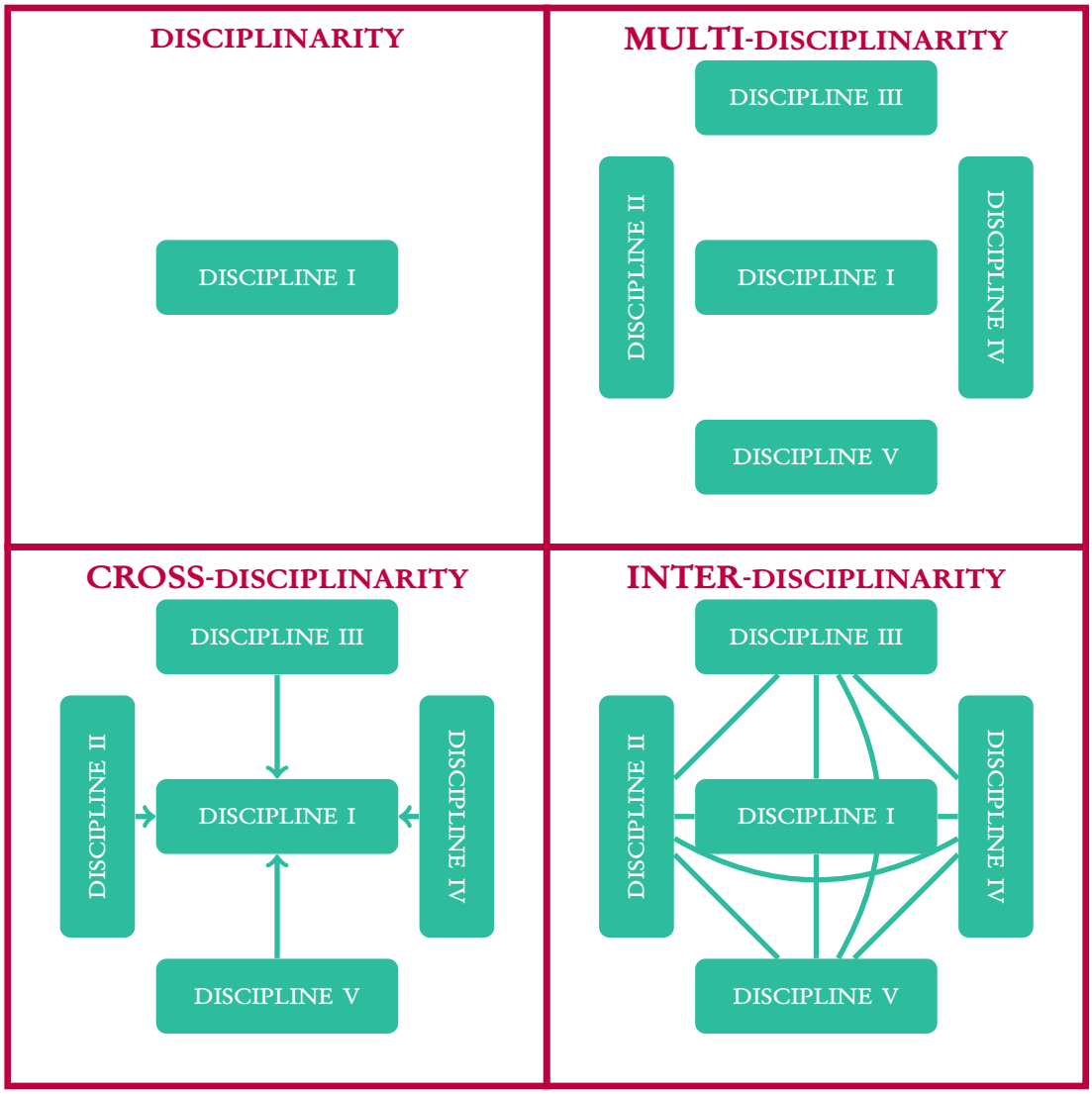


Figure 1: A graphical illustration of the differences among *disciplinarity* (DI), *multi*-disciplinarity (MULTIDI), *cross*-disciplinarity (CROSSDI), and *inter*-disciplinarity (ID). It is based on Figure 2 on p.15 of Jantsch 1970. The existence of an edge between two disciplines notates that those disciplines collaborate towards specific *ends*. In the case that the edge is *directed*, it means that the discipline at the tail of the arrow (*dominated discipline*) works towards ends of the discipline at the head of the arrow (*dominating discipline*). As *meta*-disciplinarity (METADI), I construe a practice where disciplinary experts have to go "*meta*" their discipline's boundaries. Ergo, both CROSSDI and ID are examples of METADI.

Consequently, as Jantsch 1972 remarks, ID practitioners work toward ends of *two levels*: the first lower level is about ends inside each discipline that are prerequisites for achieving the higher second-level ends. The first-level ends are *disciplinary* (henceforth *DI*) and the second-level ends are *meta-disciplinary ends* (henceforth METADI) (Jantsch 1970, pp.15-17; 1972, p.105,106).[13] In the case of CROSSDI, we have *only one* level of ends (*ibid.*). The ends of both dominating and dominated disciplines are ends *inside* those disciplines. I will call the disciplinary ends of the dominating discipline that constitute the purpose of the CROSSDI research *CROSSDI ends* (e.g., the example of defending a client in the previous paragraph is such a CROSSDIend). The practitioners of the dominating discipline may work towards non-CROSSDI ends which are prerequisites of achieving the CROSSDI ends. E.g., if the CROSSDI end is to win a trial, legal representatives will have to prepare their defense. That

[13]The abbreviation "*ID*" is taken from Schmidt 2022; Arnold 2020. "*CROSSDI*", "*MULTIDI*", and the like are abbreviations introduced by me based on the ID abbreviation. Note that Jantsch 1970, 1972 does not use the term "*meta-disciplinarity*". This is again a term that I introduce.

is a prerequisite for achieving the CROSSDI end but not a CROSSDI end itself. Finally, CROSSDI ends are DI ends *relative* to the disciplinary experts of the dominating discipline and METADI ends *relative* to the experts of the dominated disciplines: the latter have to go "*meta*" their discipline's boundaries to realise the CROSSDI end. Considering this, by "*METADI*" I will refer to both ID and CROSSDI unless specified otherwise.

So far, I have construed all disciplinarities (meta-, multi,- cross-, inter-) as *practices*. This construal of interdisciplinarity as a *practice* diverges from Jantsch's construal of interdisciplinarity as an "*organisational principle*", the organisational principle that organises the research practice among different disciplinary experts (Jantsch 1970, pp.16,18; *cf.* Jantsch 1972, p.100). My position is that Jantsch conflates the principles that guide a practice with the practice itself; interdisciplinarity as research practice is organised based on certain principles (for examples of such principles *see* §III.3), but it is not a principle itself. The construal of interdisciplinarity as a practice is also on par with Springer's 2020 *Encynclopedia of creativity, invention, innovation and entrepreneurship* entry on *interdisciplinarity* (Arnold 2020, p.1445). Notwithstanding, in the same entry, it is noted that apart from this contrual, "*interdisciplinary*" can also be used to notate "*the integration of different concepts, methods, and data*", the concepts, methods and, data of the cooperating disciplines (*ibid.*). This is a typical case of the so-called "*process/product ambiguity*", or *practice/product conflation* as I think is the more accurate in this case (*see* Vaquero 2013, ¶8): cooperative practice among the disciplines is the *process* that produces the *product* of integrated "*concepts, methods, and data*". The foregoing disambiguation is needed for answering the one-and-the-many question. To answer it, we need to clarify *what* are the many and *what* is the one. And the answer is that the many are the cooperating disciplinary *practices* and the one is the ID (or CrossDI) *practice* as a whole.

Not a principle, not a product, but a *practice*.

What is then the gluon that binds "*the many*" so as to form "*the one*" according to Jantsch's response? *Prima facie*, it seems that the gluon is the METADI common ends. However, those ends are what *motivates* the disciplines to form "*the one*" and *not* what merges them. The answer is given by Jantsch's shrewd move to distinguish between a mere *cooperation* among disciplines and a *coordinated* cooperation, coordinated towards realising the METADI ends (1970, p.14; *cf.* 1972, p.105). Therefore, the gluon is also a *practice*: the practice of *coordination*, of coordinating the cooperation of the involved disciplines. Henceforth, by "*collaboration*" I will mean *coordinated cooperation*.

gluing: the practice of *coordinating*

Let's see a specific example of how coordination can work as gluon. We saw that universities are traditionally MULTIDI departments with different disciplines like psychology, mathematics, law, chemistry, electrical engineering, etc, each pursuing its own disciplinary goals without attempting any cooperation with the rest. However, many of the institutionalised disciplines, usually the ones grouped together as departments of a specific faculty, *do* share common ends. E.g., one could argue that a goal of a faculty of natural science disciplines (e.g., physics, chemistry, planetary geoscience) is to attempt to answer the question of *how does the physical world work* (Jantsch 1970, p.19). That does not mean though that those departments cooperate *coordinated* towards the realisation of those common ends. They can still share knowledge, resources, and expertise as part of their cooperation, but that is a different thing. This *cooperation without coordination* makes the disciplines more glued than a mere *multi*-disciplinarity but less glued than both *cross-* and *inter*-disciplinarity. It is the case of *pluri*-disciplinarity (*pluriDI*) (*ibid.*, p.15).

cooperation *without* coordination

Summing up, cooperation is a *necessary* condition to practice at a *meta*-disciplinary level. However, albeit necessary, cooperation is *not sufficient* neither for *cross-* nor for *inter*-disciplinary practice. We also need *coordination*. This coordination takes place in at least three *interfaces* among the involved disciplines: *semantic*, *pragmatic*, and *teleological* interfaces (*cf.* with the concept of *trading zones* in Arnold 2020, pp.1448-1449). More precisely, since the METADI ends are *outside* the borders of the involved disciplines, the disciplinary experts have to communicate their disciplinary knowledge outside of their discipline. Using this knowledge, they also have to create new METADI knowledge. Hence, they have to produce an interface of *semantics* that will allow them to articulate and share METADI knowledge. The experts further need to communicate *what they do* with those semantics as well as compose new ways of using them. Hence, they have to produce an interface of *pragmatics*. Finally, experts have to delineate the *ends* towards which their practice is oriented, both final ends as well as prerequisite intermediate ends. Hence, experts have to compose a *teleological* interface. The *coordinated cooperation* towards the synthesis of new semantics, pragmatics, and teloi is what I call *collaboration*. Both ID and CROSSDI are collaborations and their *gluon* is the *practice* of *coordinating* their interfaces.

semantic, *pragmatic*, & *teleological* interfaces

collaboration

Legal ALGOAI is a CROSSDI practice since the *end* is to engineer an AI model that can co-produce judicial power. Ergo, that model should be able to use *legal semantics* and *pragmatics*. Therefore, all three interfaces, semantic, pragmatic, and teleological, are *dominated* by *legal science* (dominating discipline). The proposed methodology of modeling judicial justifications in CHAPTER III will mainly be about the coordination of *concepts* in the semantic interface (let's call this subset of the semantic interface as the *conceptual interface*) and the coordination of *methodologies* in the pragmatic interface (let's call this subset of the semantic interface as the *methodological interface*).

legal ALGOAI engineering: CROSSDI *not* ID

Now the million dollar question is *how* can we establish such interfaces, i.e. *how* can we apply the gluon.

This is where Isaiah Berlin's zoomorphic analogy becomes handy.

II.2.2 Foxes as gluons

Since the gluon is a *practice*, it depends on the *competence* of the experts that perform that practice. To be able to glue together semantics, pragmatics, and teloi from different disciplines, the experts should have at least some basic knowledge about the disciplines they are gluing. However, by investing time in the METADI practice of gluing, the DI expert inevitably has less time to invest in their DI research. Since each expert has a limited capital of resources to invest in the METADI research (e.g., depth of expertise, time, mental fatigue, etc), the more of that capital is invested in one practice (e.g., in DI practice), the less of that capital will be invested in another practice (e.g., in gluing).

A solution to that can be to remove the burden of gluing from the DI experts and reallocate it to another group of experts whose main task is exactly that: *to glue*. "*... at least one of the research participants has to think interdisciplinarily, working deliberately on the integration of the different methods and research findings (Parthey 1999).*" (Arnold 2020, p.1450). Those are not *disciplinary*, but *meta*-disciplinary experts (or simply the METAS). This responsibility reallocation will allow DI experts to invest more resources in their DI research and hence bring more adequate DI knowledge to the table. It will also allow the experts that perform the gluing to invest more resources both in refining the METADI interfaces and evaluate the overall METADI output. In other words, the proposed way of coordinating the cooperatarive practice among the disciplinary experts is to split the experts into two groups: (α) *prickles of hedgehogs* that focus on DI research; (β) *skulks of foxes* (i.e., the METAS) that focus on the METADI practice. Hedgehogs know "*one big thing*": their discipline. Foxes' knowledge spans across all collaborative disciplines and it goes deep enough to allow them to coordinate prickles' practices. Note that as already mentioned in §1, ¶1, an expert can be both a hedgehog and a fox. And indeed, for the foxes to glue, they will have to meet the hedgehogs in the interfaces among the disciplines. Consequently, the more "*foxness*" a hedgehog has and the more "*hedgehogness*" a fox has, the more adequate their collaboration in the METADI interfaces will be.

So far, we have seen that ALGOAI engineering teams should comprise by at least three teams of experts (§I.2.5, ¶9; §I.4): AI engineers, legitimacy experts (e.g., legal, social, & political scientists), and formal philosophers & logicians. Since I have focused on *legal* ALGOAI, for reasons of convenience, from all legitimacy experts, I will be concerned only with *legal* scientists. As I will argue in detail in §4, it is *logicians & formal philosophers* that should play the role of fox. Long story short, we saw in §I.4 that it is logicians & formal philosophers should identify & formalise judicial reaosning methods so as to incorporate them in ALGOAI models. It is those formalisations of judicial reaosning that will *glue* together the semantics & pragmatics of AI engineering & law so as to engineer legitimate legal ALGOAI models.

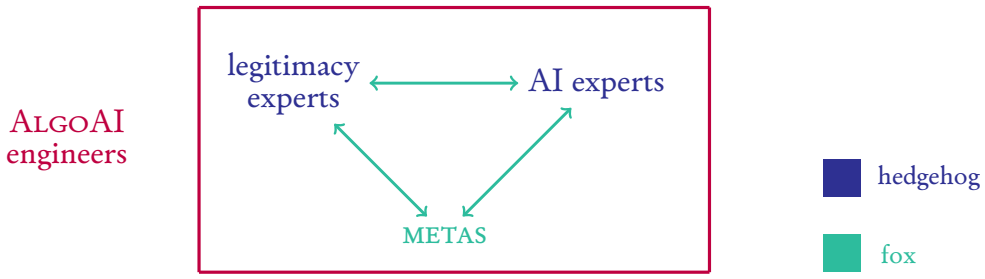


Figure 2: As *ALGOAI engineers*, I characterise the totality of experts involved in the designing of legal AI. Those are the legal & AI experts, the hedgehogs of the group, and the METAS, the foxes.

Before introducing how *logic* can glue the two prickles of hedgehogs, I would like to contextualise the endeavour of engineering legitimate ALGOAI in the context of the novel philosophical discipline of *philosophy of interdisciplinarity* (henceforth *PhID*).[14] I do so in order to explicate two core aspects of ALGOAI engineering: (α) its *trans*-disciplinarity. I.e., the collaboration among disciplines so as to produce knowledge & ontology that *transcends* academic disciplines to realise non-academic oriented ends (political, legal, social, etc); (β) the nuances & dangers that lurk during the production of *contactual information* by the collaboration of the different disciplines.

[14] Abbreviation borrowed from Mäki 2016.

II.3 Philosophy of interdisciplinarity

II.3.1 Philosophy of science's nascent evolutionary stage

So far, I have raised questions regarding the merging of scientific theories, methodologies, concepts, and practices, regarding the boundaries between disciplines and how those boundaries differ depending on the particularities of different scientific research projects, regarding the combination of disciplines to introduce new ID disciplines (e.g., forensic science), regarding the co-operation of experts from diverse disciplines to develop new METADI objects (e.g., AI models of cancer detection). I have further raised the question of what is a science in the first place and I have taken specific positions about whether evaluative judgements can be scientified or not as well as whether we can acquire new knowledge about an ontology without sufficient *prior* epistemic access to that ontology. It becomes evident that there is a robust relation between *philosophy of interdisciplinarity* and *philosophy of science* (*PhiSci*).[15] The reality is that PhID does not merely overlap with PhiSci. PhID *is* PhiSci.

In their 2016 PhID manifesto "*Philosophy of interdisciplinarity: What? Why? How?*", Uskali Mäki argues that PhID is the current third emergent stage in the evolution of PhiSci. They further provide the following streamlined historical account of the three stages of the evolution of PhiSci (the stages emerged in the provided order). Note that in contrast to human evolution, antecedent PhiSci evolutionary stages do not cease to exist, but they continue to spawn fruitfully (Mäki 2016, pp.8-9):

GENERAL PHILOSOPHY OF SCIENCE deals with science as a whole, asking question like how should a scientific methodology or a scientific explanation should look like *across* sciences. Such is for instance the case for the *Classical model of Science* (*CMS*), a normative model of how scientific explanation should be *in general*, a model that originates from Aristotle's general philosophy of science (de Jong and Betti 2010) and that can be specialised to model explanations in legal ALGOAI (Iatrou 2022a). Apart from Aristotle, other important pioneers of general PhiSci the Enlightenment's precursors are Francis Bacon (1561–1626) and René Descartes (1596–1650), and Enlightenment's Gottfried Wilhelm Leibniz (1646–1716), Immanuel Kant (1724–1804), Isaac Newton (1642–1727), Galileo (1564–1642), Antoine-Laurent Lavoisier (1743–1794).[16]

PHILOSOPHIES OF SPECIALISED SCIENCE like philosophy of computer science, physics, psychology, law, international relations, forensic and cognitive sciences, as well as AI.[17] Even if some of the aforementioned examples are not considered by some as *sciences*, such a debate about what is science is *still* part of PhilSci (Sven Ove Hansson 2021). In this evolutionary stage, scientific practice is identified as *disciplinary practice* with each discipline-science developing its own semantic, pragmatic, and teleological aspects. The general remarks of the *general PhilSci* from the previous evolutionary stage are adjusted to fit the particularities of each discipline like the CMS being adjusted to fit the discipline of law (Iatrou 2022a, §§2.3-2.4).

PHILOSOPHY OF INTERDISCIPLINARITY: is about the way disciplines interact at a *meta*-level to compose new scientific practices: new semantics, new pragmatics, new ends, including *exo*-scientific ends that one does not encounter in the traditional Enlightenment conception of science. While the ideal for a scientific practice was to separate facts from values, ID wants to bring *values* back into the picture so as to solve "*real-world*" problems like engineering ALGOAI that meets specific legitimacy requirements (*cf.* §3.1.1; *see also* Schmidt 2022, pp.95-96; *cf.* Reiss and Sprenger 2020, §3; Capaldi 1998, pp.,13-14,295).

Before moving forward, allow me to make a few important clarifications. A scientist of an ID science (e.g., forensic science, cognitive science, AI, or decision analysis) is still a scientist of a specialised science. In other words, from the moment an ID discipline is formed, it *is* a discipline. Hence, philosophy *about* that discipline is still a philosophy of a specialised science belonging to the second evolutionary stage of PhiSci. However, whenever one reflects on how semantic, pragmatic, and teleological aspects from different specialised sciences can merge to form a new discipline, then one engages in a philosophical discourse about what happens at a *meta* level of those disciplines. It is no longer a philosophy about a specialised science, but it is a philosophy about what happens in the *meta*-level of specialised sciences. Ergo, *philosophy of meta-disciplinarity* (or *PhiMetaDI*) would be a more appropriate name for this new philosophical discipline with philosophy of interdisciplinarity being a sub-discipline. Having said that, I will stick to "*philosophy of interdisciplinarity*" since it has already

[15] Abbreviation borrowed adjusted from Russo 2022.

[16] Due to the fuzzy borders between science and philosophy at the time, many of the aforementioned (Newton, Galileo, Lavoisier) are also considered *scientists*. They are usually labeled under the umbrella term *natural philosophers* (Kitcher 2023).

[17] *See respectively* Angius, Primiero, and Turner 2021; Frisch 2022; Robins, Symons, and Calvo 2020; Leiter and Sevel 2022; Joseph and Wight 2010; Meester and Slooten 2021; Isaac 2020; Boden 1990.

dominated the literature.

Based on the foregoing, two central topics in the PhID are: (α) *transdisciplinary* (henceforth *TRANSDI*) *ends*. I.e., ends that *transcend* academia orienting disciplinary practices towards "*real-world*" problems; (β) *gluing* specialised disciplines so as to produce what is called *contractual information*. In what follows, I elaborate more on those two topics that will be of relevance later on.

III.3.1.1 *Trans*-disciplinarity: erecting legitimacy pillars

The goal of engineering an AI model that exercises political power *legitimately* is a *political* and a *societal* goal. I.e., it is a goal that *transcends* the disicplinary practices of academia, and hence, it is what is called a *transdisciplinary* (TRANSDI) goal (Jantsch 1972, pp.16-17). METADI practices like ALGOAI engineering whose purpose is to realise TRANSDI ends are what Schmidt 2022 (pp.29-32) calls TRANSDI-*oriented* practices. The reorientation of academia in the 70's towards the resolution of TRANSDI problems was what motivated the emergence of all those concepts of X-disciplinarities (*cross*-disciplinarity, *inter*-disciplinarity, *plural*-disicplinarity, etc) from top-down unelected bodies like the OECD.[18] As argued in *fn.* 11, what established Jantsch's 1972 paper as one of the foundational papers about the contemporary conception of interdisicplinarity was its submission to the 1972 OECD Paris conference "*Interdisciplinarity: Problems of teaching and research in universities*", a conference whose purpose was not only to promote interdisciplinarity, but to promote interdisciplinarity as the appropriate means to deal with complex TRANSDI problems. It is not accident that it was co-organised by the OECD's organ CERI (*Centre for Educational Research and Innovation*) whose objectives are *inter alia* to support the OECD member states and their partners in gluing educational research with policy development (Apostel et al. 1972, p.4; OECD's Directorate for education and skills 2021.). Even the precursors of the 1972 cocnept of interdisciplinarity were motivated by TRANSDI-*oriented* goals: "*The modern use of the term "interdisciplinary" goes back at least to the 1940s, when scientists as well as newly found private funding institutions in the United States, like the Carnegie Corporation as well as the Rockefeller and the Ford Foundation, tried to encourage innovative research beyond traditional disciplinary boundaries ([Jamie 2014, pp.76-103; Harvey J. 2015])*" (Arnold 2020, p.1445). ALGOAI engineering is essentially a manifestation of this TRANSDI reorientation of academia. Note that this distinction between academic/TRANSDI ends is essentially the premiss of the new separation of powers. As argued in §I.2.5, the new separation of powers is predicated on the distinction between academic (*factual*) and TRANSDI (*value*-ladden) problems: epistemic authorities should follow the TRANSDI value-laden decisions of the traditional authorities and restrict their activity to more factual judgements.

*trans*ending academia

In order to realise the TRANSDI ends of their practice, ALGOAI need to first *produce* information by merging the semantics & pragmatics of their disciplines in a METADI level. I.e., they need to produce what Mäki 2016 names *contactual information*.

II.3.1.2 Contactual information

Mäki 2016 (p.10) argues that the philosophical discourse in PhID produces two types of information: (a) *comparative information*. I.e., information about similarities and differences among disciplines; (b) *contactual information*. I.e., the output of gluing disciplines together.[19] I do embrace Mäki's position as long as comparative information is subjugated by contactual information: comparative information is produced with the purpose of producing contactual information. For instance, in CHAPTER IV, I *compare* the concept of *causal explanation* in the disciplines of law and logic so as to *glue* them, i.e., so as to produce glued *contactual* information. I make this restriction to comparative information since comparative information has traditionally been used in both general PhiSci (e.g., comparing different types of explanation among disciplines to abstract a higher level model of scientific explanation like the CMS) and specialised PhiSci (e.g., comparing the concept of causation in law with causation in natural sciences so as to provide a better understanding of the former like in Hart and Honoré 1985, pp.11-12). Ergo, it would be historically (and conceptually) inaccurate to classify comparative information only to the third evolutionary stage of PhiSci. After all, Mäki themselves acknowledge that conceptualising PhID to include any type of comparative information is "*a rather broad conception of what the philosophy*

comparative *v.* contractual information

[18] *Organisation for Economic Co-operation and Development* (*OECD*). It is another international organisation whose precursors were founded right after WWII to engineer a European order based on similar to the CoE legitimacy requirements like revitalising and reshaping the European economy based on the value of *democracy*. Today, it has transcended the European borders reaching a global range with a total of 38 member states 26 of whom are also CoE member states (https://www.oecd.org/about/members-and-partners/, accessed 25 March, 2023).

[19] Regarding contactual information, Mäki (*ibid.*) explicitly differentiates between "***mere combinations*** *of bodies of disciplinary information*" (emphasis added) and *glued* METADI information: "*The production of contactual information requires going beyond mere combinations of bodies of disciplinary information generated by philosophies of special sciences. One must analyze the large variety of ways in which disciplines can be in consequential contact with one another – such as collaboration, inspiration, transfer of models or methods, evidential support or criticism, integration and unification, and so on.*".

of interdisciplinarity would cover" (Mäki 2016, p.10).

In what follows, I firstly provide a *typology* of contactual information that can be useful to identify strategies to deal with problems about producing contactual information: different types of contactual information will face different problems which will require different solution strategies. Afterwards, I provide an example of such a challenge for legal ALGOAI engineers, one of the most notorious controversial challenges in the AI & Law discipline that has become known as the *hungry judge effect*.

II.3.1.2.1 A typology of contactual information

Depending on the *type* of contactual information, METADI practice can be classified into different overlapping types. At least three of those types are of relevance forALGOAI engineering.[20] This tripartite typology is premissed on three central concepts of PhiSci: (α) *object*; (β) *theory*; (γ) *methodology*. The respective types of METADI are: (α') *object*-oriented METADI; (β') *theory*-oriented METADI; (γ') *methodology*-oriented METADI.

• *Object*-oriented METADI is either about *understanding* objects (e.g., the human brain, the ozone hole, nanoparticles, the stock market, causal inference in law, ice melt, loss of biodiversity (Steffen et al. 2005, p.91; *cf.* Schmidt 2022, p.93)) or *constructing* new ones intentionally (e.g., cyber-physical space, AI, nuclear power plants, skyscrapers, water supply systems, international relations systems, military infrastructures, virtual reality (VR)).[21] Note that such constructed METADI objects are constitutive elements of SOCIETY 5.0.

Now *object*-oriented METADI is premissed on *ontological non-reductionism*: the ontologies of individual disciplines do not suffice by themselves to understand and/or create the METADI object (p.27). An insightful way of construing the position of ontological non-reductionism is by using the construal of interdisciplinarity found in Heckhausen 1972 (pp.80-81), another paper published in the proceedings of the 1972 OECD conference.[11] According to Heckhausen 1972, the collection of *objects* that constitute the object of inquiry of each discipline (what Heckhausen calls *material field*) are different from the way that each discipline *views* those objects in terms of semantics and pragmatics (what Heckhausen calls *subject matter*). E.g., the human brain belongs to the material field of neurology, cognitive science, forensic science, biology, psychology, and so forth, but each of those disciplines has its own semantic and pragmatic view of the human brain, i.e., its own distinct subject matter. Ontological non-reductionism is when we have to combine the subject matters of multiple disciplines so as to understand/construct objects in their material fields.

matterial field *v.* subject matter

• *Methodology*-oriented METADI is premissed on *methodological non-reductionism*: disciplinary methodologies are not sufficient by themselves to realise METADI ends. Questions of methodology-oriented METADI that will be addressed in CHAPTER III are *how to evaluate* METADI practice's outcomes, *how to transfer* knowledge among disciplines, *how to compose* METADI methodologies, as well as *how to produce* knowledge and ontology that satisfies TRANSDI ends (Schmidt 2022, p.28).[22]

• *Theory*-oriented METADI is premissed on *theoretic non-reductionism*:[23] disciplinary theories are not sufficient by themselves to realise METADI ends. E.g., we will see in CHAPTER IV that formal theories of causal inference (e.g., from the disciplines of logic, probability theory, and logical programming) are incapable of modelling causal justifications in the ECtHR case-law. This is where other disciplines like jurisprudence and human rights law come to the rescue. Questions of *theory*-oriented ID that will be addressed in CHAPTER III are *how to compose* METADI theories, models, laws, and explanations and *whether* the proposed theories, models, laws, and explanations provide an *adequate understanding* of the objects of the METADI material field (Schmidt 2022, pp.27-28).

[20] This 3-typed typology is based on a similar proposal made by Schmidt 2022 (pp.26-35), albeit there are certain differences that I will highlight.

[21] Schmidt does not make this explicit distinction between *understanding* and *constructing* an object (2022, p.27).

[22] Note that Schmidt uses "*method-oriented*" instead of "*methodology-oriented*" (*ibid.*). Personally, I adopt Howell's 2013 disambiguation of *method* and *methodology* according to which *methodology* is the broader research strategy (e.g., causal justification) and *method* a particular "*means*" or "*modes of data collection*" that comprise the research strategy (e.g., a specific type of causal justification like the but-for test that will be explained in CHAPTER IV) (*ibid.*, pp.xi-x). Based on this disambiguation, I deem more appropriate the term "*methodology-oriented*" since what Schmidt characterises as *method-oriented interdisciplinarity* does not concern only particular methods, but also the broader strategies that those methods are employed to achieve.

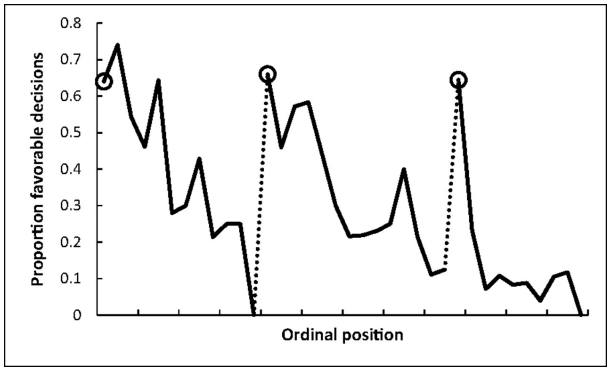
[23] Schmidt 2022 (p.28) uses "*epistemological non-reductionism*" instead which I found to be rather broad. I also deem epistemological non-reductionism to be more appropriate for methodological non-reductionism than theoretic non-reduciotnism (albeit again it is a rather broad term to describe methodological non-reductionism) since a methodology is about *how* the user of that methodology *access knowledge*, while theories can explain aspects of reality without providing an adequate account of *how* we can *know* those aspects (de Jong and Betti 2010, p.201). E.g., Platonism explains truths about mathematical propositions without providing an adequate answer of how we can access the Platonic realm of mathematical objects (Benacerraf 1983). Note that Schmidt 2022 uses neither "*methodology non-reductionism*" in his construal of *method*-oriented interdisciplinarity. *Cf.* §IV.2, ¶3

Prima facie, it seems that ALGOAI engineering is an *object*-oriented METADI practice: the engineers want to construct an object, the ALGOAI model. *Contra* Schmidt 2022 though, my position is that in different stages of the same METADI practice, the experts *alternate* between the three orientations. For instance, when ALGOAI engineers collaborate to construct the ALGOAI model (*object*-oriented METADI), they have to combine methodologies from their disciplines so as to construct the model like using logical programming to model judicial causal inference methods (*methodology*-oriented practice). At the same time, they have to glue disciplinary theories in order to justify their choices like justifying why a specific type of judicial causal inference is compatible with a specific type of logical programming and why their combination is suitable for constructing the desired ALGOAI model (*theory*-oriented practice).

In CHAPTERS III & IV, I focus on the *methodology*-oriented practice of ALGOAI engineering. More precisely, I propose a methodology that can be used to engineer parts of the ALGOAI model, with an emphasis on what is the role of logicians & formal philosophers as foxes during this *methodology*-oriented practice. Still, even in that case, we will see that it is inevitable not to alternate among the three dimensions. Subsequently, an adequate METADI *methodology* should also account for adequate METADI *theories* and *objects*.

In what follows, I provide an example of METADI practice that stresses the necessity of fleshing out full-fledged philosophical accounts of how *methodology*-oriented practices like methodology *transferring* and methodology *gluing* should be like. That example also showcases the impact that suboptimal methodology-oriented practices can have to the realisation of TRANSDI ends.

II.3.1.2.2 No justice without breakfast



source: Danziger, Levav, and Avnaim-Pesso 2011, Figure 1

The **x axis (ordinal position)** represents the time of the day in which a judgement was delivered. The **dashed lines** represent the times of the day in which the authorised experts took a meal break. The time periods between the dashed lines are the **judgement sessions**. The **circled points** represent the first judgement delivered in each of the three judgement sessions. Because unequal session lengths resulted in a low number of cases for some of the later ordinal positions, the graph is based on the first 95% of the data from each session.

The **y axis (proportion of rulings in favor of the prisoners)** represents the percentage of judgements against the status quo (and ergo in favour of the prisoners) delivered at a particular time x_i for all 50 days in the 10-month period. E.g., approximately 65% of all 1st judgments for every 1st session for all 50 days (i.e., the first circled point) were against the status quo and approximately 35% were in favour of the status quo.

In 2011, one of the most influential articles in the practice of modelling judicial judgements was published in the Proceedings of US' National Academy of Sciences (PNAS): by Danziger, Levav, and Avnaim-Pesso. The conclusion of the authors was the type of sensational thumbnail caption that pop scientific media crave for: "*merely taking a food break—may lead a judge to rule differently*" (p.6892). A position that became known as the *hungry judge effect* (Chatziathanasiou 2022, p.452).

Danziger, Levav, and Avnaim-Pesso's 2011 research is not part of the AI & Law discipline, but it has had substantial impact both in that discipline as well as in the policy-making discourse about accommodating legitimate legal ALGOAI (Chatziathanasiou 2022, §B and pp.462-463). Regarding actual policy-making, the hungry judge effect and similar studies of judicial bias like those we saw in §I.3.1, ¶1 have been used to argue in favour of accommodating legal ALGOAI that reduces judicial biases. The argument is pretty straightforward: AI does not get hungry and ergo it can be more *just* than human judges. To generalise the argument, an argument in favour of replacement legal ALGOAI is that since we already consider as *legitimate* court decisions from judges that inevitably make non-rational judgements at some point due to their human nature, we should also accept as *legitimate*, ALGOAI that makes the *same if not fewer* non-rational judgements (Chatziathanasiou 2022, p.455). Those are essentially more arguments in favour of the epistocratic post-Enlightenment position (§I.2.8, ¶4).

Let's get into the particularities of the research. 8 judges presided over parole boards (the judges were Jewish-Israeli with 2 of them being female) that serve 4 Israeli prisons delivering judgements that either grant the prisoner's request (ergo against the status quo) or reject/postpone the judgement (in favour of the status quo) for 50 days over a 10-month period. Each working day was divided into three sessions with two meal-breaks in-between the sessions. The prisoners were convicted felons (i.e., their misdeeds were of high seriousness like embezzlement, rape, or murder) and the requests were either about being granted parole or about changing the terms of an already granted parole (e.g., removing a tracking device) or about changing terms of incarceration

(e.g., prison relocation). Danziger, Levav, and Avnaim-Pesso's dataset consisted of 1.112 judgements. The method of generalising from the sample to the whole population was logistic regression. Danziger, Levav, and Avnaim-Pesso balanced those generalisations across ethnicity (Arab, Jewish), gender (male, female), availability of rehabilitation program should the prisoner be granted parole (Yes/No binary answer), gravity of the offense (scale from 1 to 7), number of previous incarcerations, and number of months served in prison.

Chatziathanasiou 2022 makes a good case as to how the results of the Danziger, Levav, and Avnaim-Pesso 2011 research fall short in *gluing* adequately the involved disciplines resulting in ill-founded conclusions. A majority of those mistakes are *methodological* mistakes including the *transferring* of methodologies. For instance, the methods the authors used to identify biases constitute research methodologies of *experimental psychology*. A criticism they received from experiment psychologists was that in general psychological effects reported for the first time tend to be larger than in later replications and consequently, they should have made a more conservative interpretation of their results. Indeed, as one can see in Figure 1 of Danziger, Levav, and Avnaim-Pesso 2011 (*see* the Figure above), the psychological effect recorded is unreasonably high: 100% (!) of the judgements delivered at the end of the first and third sessions were against the interest of the prisoners. As psychologist Daniel Lakens argued on his blog, an experimental psychologist at the Human-Technology Interaction group at Eindhoven University of Technology, "*If hunger had an effect on our mental resources of this magnitude, our society would fall into minor chaos every day at 11:45.*" (Lakens 2017).[24]

Since such types of studies are used to realise TRANSDI ends, mistakes on their methodology are not mere problems that are solved behind academia's closed doors. Such mistakes have an impact in actual *policy-making*, in the way that we engineer our social order based on their outcomes. It is this type of sloppy arguments that are carelessly used to minimise the importance of ALGOAI's bias legitimising an illegitimate exercise of power (§3.1.2.2, ¶2). This is why authors like Chatziathanasiou 2022 advocate for more clear delineation of *how* experts should cross disciplinary boundaries especially when dealing with questions that structure our lives.

II.4 Assembling Team Rocket

So far, I have argued that Kissenger's "*eminent thinkers*" when it comes to legal AI are two prickles of hedgehogs (the AI engineers and the legitimacy experts) and one skulk of foxes (the METAS who include formal philosophers & logicians). I have also argued that their collaborative practice is a CROSSDI practice. In CROSSDI, the prickles of hedgehogs of the dominating disciplines are the *leaders*: they provide guidance on *which* CROSSDI ends shall be pursued, *how* they should be pursued, how their results can be *evaluated*, and so forth. Ergo, those hedgehogs enjoy *royal* status. The experts of the rest of the disciplines are the skilled *knights* that execute the decrees of the royal hedgehogs, without that precluding their research autonomy.

Considering the above, I will introduce some basic information about the two prickles of hedgehogs that will be of use, and then, I will show how their relation to *logic* has (not) been used to *glue* them together in the discipline of AI & Law. Note that my account of the hedgehog is neither exhaustive nor rigid. That would be impossible after all since disciplines and their boundaries are fluid and continuously evolving, with the *boundary* issue being one of PhID's central challenges (Schmidt 2022, pp.24-26; *cf.* Mäki 2016, pp.5-6).

II.4.1 Two prickles of hedgehogs

II.4.1.1 The knightly hedgehogs: *AI engineers*

I have already introduced the concept of AI and its relation to logic in §I.3.2.1.1. In this subsection, I will further provide information about the concept of *engineering* that plays a pivotal role in at least three central topics of this Thesis: (α) AI engineering; (β) Enlightenment's mechanistic conception of the world; (γ) conceptual (re)-engineering (*see* CHAPTER III). More precisely, according to Chalmers's 2020 overview of conceptual (re-)engineering, *engineering* is the process of *designing*, *building*, and *analysing* an object (p.2). In AI, that object is an AI *model* (more on *what is a model* in §III.1). What does it mean though to *design*, *build*, and *analyse* a model for the ALGOAI engineers? To explicate those three concepts, I will disentangle the usual confusion among the concept of an *algorithm* $\mathcal{A}$, a *programme* Π in which $\mathcal{A}$ is expressed, the *language* $\mathcal{L}$ in which Π is written, and the *machine* $\mathcal{M}^{\mathcal{L}}$ that implements $\mathcal{A}$ as described by Π. This distinction is borrowed from Gabbrielli and

What is *engineering*?

[24] It is no accident that I cited a blog entry to criticise Danziger, Levav, and Avnaim-Pesso's 2011 methodology transferring attempt, an entry also used by Chatziathanasiou 2022 in a prestigious peer reviewed journal. As Chatziathanasiou argues and as we will see in §III.3.2.4, it is advisable for experts partaking in METADI research to use non-conventional means to communicate knowledge form their disciplines to experts from other disciplines. This of course entails that we should take appropriate measures to filter out suboptimal information like having multiple experts from the same discipline in the same ALGOAI team to hold each other accountable. Note that I have implemented this principle in this Thesis as well, but only for citations of secondary importance or as auxiliary citations next to traditional ones.

Martini 2010, §1. The resulting explanation can be generalised to other types of engineering like coneptual (re)-engineering.

Let's start with what is an *algorithm* $\mathcal{A}$. I construe an algorithm as a set of *instructions* that dictate how from a set of initial conditions (the *input*) we can reach to a set of desired conditions (the *output*) (*cf.* Angius, Primiero, and Turner 2021, §3). A classical example, and for some the first ever example of a non-trivial algorithm (Russell et al. 2021, p.27), is Euclid's instructions on how to calculate the greatest common divisor (GCD) between two natural numbers. In this case, the *input* is two natural numbers and the *output* is the GCD. The term "*algorithm*" originates from the name of the Persian mathematician Muhammad ibn Musa al-Khwarizmi (780–850 AD), who provided instructions for performing arithmetic operations using Arabic numerals (Angius, Primiero, and Turner 2021, §3). In the 19th century, logicians and mathematicians like George Boole started the ambitious endeavour of introducing algorithms for performing logical deduction so as to do mathematics. It was those endeavours that set the foundations for 20th century's logic-based AI (Russell et al. 2021, p.27).

What is an *algorithm*?

We saw that each algorithm $\mathcal{A}$ consists of a set of instructions $\mathcal{A}$. Now the same set of instructions can be written in multiple different languages $\mathcal{L}_i$. Euclid wrote his algorithm in ancient Greek while Greek high school students write it in modern Greek. Even in the same language, the same algorithm can be expressed differently. A specific expression of an algorithm $\mathcal{A}$ in a language $\mathcal{L}$ is a *programme* $\Pi^{\mathcal{L}}_{\mathcal{A}}$. At the same time, I do not speak Slovak so I would not be able to execute Euclid's instructions written in Slovak. In other words, a programme $\Pi^{\mathcal{L}}_{\mathcal{A}}$ is executable only if those that execute it speak the language $\mathcal{L}$. In the discipline of computer science, and hence in its subdiscipline of AI, those that execute programmes $\Pi^{\mathcal{L}}_{\mathcal{A}}$ written in a programming language $\mathcal{L}$ are the so-called *abstract machines* $\mathcal{M}^{\mathcal{L}}$. By "*abstract*", one means that the machine does not have to be *physical* like the laptop where I am writing this text, a typical case of a physical machine in computer science consisting of *logic* circuits and electronic components both quintessential advances of INDUSTRY 3.0 ([]; *cf.* §I.2.6.1).

programme *v.* algorithm

Taking into consideration the above, in the practice of engineering an object $\mathcal{O}$ (e.g., an algorithm $\mathcal{A}$), I construe the practice of *designing* $\mathcal{O}$ as the practice of using a language $\mathcal{L}_d$ to identify *components* of $\mathcal{O}$ and the *properties* of those components (e.g., identifying $\mathcal{A}$'s set of instructions and their order), as well as identifying which is the *use* of $\mathcal{O}$ (*cf.* §III.1). $\mathcal{L}_d$ is the language the engineers use to *discuss* which are those components/properties/use (*e.g.*, the ordinary language infused with terminology from computer science, mathematics and logic; *cf.* §III.1.1). Next, I construe the practice of *building* $\mathcal{O}$ as the practice of giving to $\mathcal{O}$ flesh & bones (e.g., writing a programme $\Pi^{\mathcal{L}_b}_{\mathcal{A}}$ in a language $\mathcal{L}_b$ so as to be executed by an abstract machine $\mathcal{M}^{\mathcal{L}_b}$). In §I.2.5, ¶7, the language $\mathcal{L}_1$ is the language $\mathcal{L}_d$ of designing ALGOAI models, while the language $\mathcal{L}_2$ is the language $\mathcal{L}_b$ of building the designed models. Finally, I construe the practice of *analysing* the now built $\mathcal{O}$ as the practice of *evaluating* whether $\mathcal{O}$ has the components and properties identified during the designing phase as well as whether it is adequate enough be used as intended. Evaluation is performed using another language $\mathcal{L}_e$ that can differ from $\mathcal{L}_d$. E.g., ALGOAI engineers may need to perform statistical evaluative tests like t-tests, while there is no use of statistics in the designing phase. Note that my construal of all three practices (designing, building, analysing $\mathcal{O}$) is open-ended. Concluding, as we will also see in §III.3.3, ¶1, the three practices are not always performed linearly. Sometimes they may even be performed simultaneously (§III.3.2, ¶1).

ENGINEERING: designing, building, analysing

II.4.1.2 The royal hedgehogs: *legal experts*

It is time to bite the bullet and talk about *legal* ***science***. Vaquero 2013 (§2) provides a concise account of five models of legal science at least two of which are of relevance for ALGOAI engineering. I will introduce those models based on their different approaches regarding how the law should be *interpreted & applied* by judicial authorities since the interpretation & application of the law is the type of power that is primarily responsible for potential ALGOAI misorientations (§I.3.2.1, ¶2). I will first provide a standard for the literature *logical* modelling of interpreting & applying the law,[25] and then, I will explain how those five types of legal science differ based on their use of that logical model.

To begin with, legal science can be construed as one of the disciplines whose material field is *law*. Other such disciplines are legal anthropology, history of law, judicial politics, and sociology of law (Vaquero 2013, ¶9). What differentiates legal science from the rest is *inter alia* its subject matter which is the *interpretation & application* of law by judicial authorities. Note that we should not conflate the legal scientists with the judicial authorities that perform the interpretation & application of the law (*ibid.*, ¶22; Vaquero calls them "*legal operators*"). The legal scientists are concerned with either how the law *is* interpreted & applied by judicial authorities (*descriptive* legal science or legal science *stricto sensu*) or with how judicial authorities *should* interpret & apply the law (*normative* legal science or *legal dogmatics*). Both *legal science stricto sensu* and *legal dogmatics*

LEGAL SCIENCE: descriptive *v.* normative

[25] It is a model I explicated in Iatrou 2022a (*see also* Iatrou 2022b). As we will see in the rest of the chapter, its variations have been used for more than a century by (formal) philosophers, logicians, legal scientists, as well as (GOF)AI engineers.

constitute the *legal science ampio sensu* (henceforth simply *legal science*) (*ibid.*, ¶27).[26] Two of the five models of legal science are descriptive: (α) the *normativistic* model; (β) the *realistic* model. The rest three models are about the *normative* conception of legal science: (γ) the *argumentativist* model; (δ) the *realistic-technological* model; (ε) the *critical* model.

AlgoAI engineering should alternate between *both* descriptive and normative models depending on the engineering phase. In principle, as argued in §I.3.2.1.2, ¶11, the AlgoAI engineers should engineer AlgoAI models based on how judicial authorities *do* interpret the law (*descriptive* legal science). Otherwise, they *illegitimately* substitute judicial authorities ending up engineering an illegitimate model. At the same time though, we also saw in §I.3.2.1.2, ¶11 & §I.3.3 that AlgoAI engineers should criticise, provide feedback, recommend, & check-and-balance judicial authorities by signaling to judicial authorities problematic aspects of their practice before accommodating them to AlgoAI models. In this case, the AlgoAI engineers will have to resolve to a *normative* account of which are those problematic aspects (e.g., logical contradictions or incoherence), recommend & give feedback on how they should be dealt with, etc (*normative* model of legal science). The question now becomes which of those 2 descriptive & which of those 3 normative models are suitable for the AlgoAI engineering practice? To answer the question, I will first attempt a *logical* construal of *interpretation & application* of the law.

interpreting the law

A core, if not *the* core, demarcation criterion among the five models is their different approach to the *interpretation* of the law. Interpretation can be construed as the decision of whether a particular term t is *subsumed* by a specific concept $\mathcal{C}$ (MacCormick 1992, §IV; *cf.* with the *but-for test* in §IV.2.1.2.1). I.e., whether a particular is an instance of a universal or whether the first-order logic formula $\mathcal{C}(t)$ holds in the *actual* world. I will name the criteria of whether a term is subsumed by a concept *subsumptive criteria* (or *subsumptive tests*). E.g., a subsumptive test of whether an animal t is subsumed by the concept $\mathcal{C}$:=TIGER is performing a DNA test on t's DNA. Or a subsumptive test for whether the *aim* of a state's interference to the freedom of expression is a *legitimate* aim (i.e., whether that aim t is *subsumed* by the concept $\mathcal{C}$:=LEGITIMATE) is to check whether t is listed in ¶2 of ARTICLE 10 like the legitimate aims of national security, territorial integrity, and protection of health (ECtHR Registry 2021, §III.B.2).

subsumptive test

application of the law

What about the practice of *applying* of the law? The application of the law is predicated upon the interpretation of the law. More precisely, *applying* the law can be construed as performing a particular type of *inference* called *deontic inference*. A deontic inference can be construed as an inference whose premisses are comprised of: (α) *imperatives* that *prescribe* how a collection of objects (the *domain* or *juridistiction*) should behave under specific circumstances. (β) *declaratives* that describes the *actual* conditions of a particular subset of the domain. Based on which are the *actual* conditions of those particulars, we can *deduct* which of the imperatives hold for those particulars. That is the *conclusion* of the deontic argument.[27] Ergo, *applying* the law is essentially deducting which imperatives hold for specific objects of the law's jurisdiction. To decide whether a particular object t meets a condition $\mathcal{C}$ required to apply an imperative is essentially deciding whether t is subsumed by $\mathcal{C}$. In other words, deciding whether the law is applied in a particular case depends on how we *interpret* the *legal* concepts that constitute the *legal* provisions. This process of *deducting via subsumption* is what MacCormick 1992 calls *subsumptive-deductive* inference (*cf.* Alchourrón 2015) and it can be modelled using the following *inference schema* (Iatrou 2022a):

$$\frac{\begin{array}{rl}\psi(x) \Leftarrow \varphi(x) & \text{GENERAL IMPERATIVE}\\ \varphi(\alpha) & \text{PARTICULAR DECLARATIVE}\end{array}}{\psi(\alpha) \quad \text{PARTICULAR IMPERATIVE}}$$

where φ & ψ are propositional functions, x is a variable, and a a term without free variables (the arities of x & α can vary). φ & ψ describe certain states of the world. The operator "$\Leftarrow$" is used to construct imperatives. Specifically, $\psi(x) \Leftarrow \varphi(x)$ notates that *if* the conditions described by the universal φ are met, then the conditions described by the universal ψ *must* or *can* follow. I say "*can*" and "*must*" since the traditional modalities that discern *imperatives* from *decleratives* are the modalities of *permissibility* (what *can* be the case) and *obligation* (what *must* be the case), the so-called *deontic modalities* (Palmer 2001, §1.3.2). The reader familiar with deontic logic may be perplexed as to why I did not use the traditional deontic modal operators of

[26] "*stricto sensu*" and "*ampio sensu*" mean "*in a strict sense*" and "*in a broader sense*" respectively. In other words, Vaquero 2013 attributes to scientific practice in the *strict sense* the property of *descriptiveness*; science argues how the world *is* (e.g., how nature *is*), *not* how the world *ought to* be. Ergo, a normative account of how law *should* be interpreted and applied is a *stretched* definition of legal science.

[27] For more on what should be construed as an *imperative* and what as a *declarative see* Hilpinen and McNamara 2021, §4. For the Thesis, the intuition that imperatives prescribe what *should be* the case while declaratives describe what *actually is* the case suffices (*cf.* §I.1, *nn.* 1 & 2; §I.2.3, ¶7).

permissibility and obligation used in standard deontic logic (P and O respectively)[28] and instead I preferred a *conditional* ($\Leftarrow$). The answer is that in traditional logic-based legal AI (i.e., in legal expert systems), it is common practice to use simple IF-THEN rules (*fn.* 79; *cf.* §.4.2.1) like the rule in Example 2 below. In such rules, the THEN-clause can be interpreted by the user as a deontic modality clause. I.e., while the object language may not *per se* have modalities, the *meta*-language of those that *semantically* interpret the object language *does* (*cf.* §I.3.2.1.1, ¶9). The logician may also object to the fact that while I write that the conclusion of the subsumptive-deductive inference schema above is an *imperative*, its logical form does not contain the operator "$\Leftarrow$" which is used to construct imperatives. I did so because I wanted to show how the deductive-subsumptive schema is used in expert systems to acquire new knowledge. More precisely, traditionally, legal expert systems are comprised of a conjunction of imperatives like $\psi(x) \Leftarrow \varphi(x)$ that model actual legal provisions. The user of those models, inputs facts of a case (i.e., particular declaratives like $\varphi(\alpha)$), and then, the imperatives which have the inputted facts in the IF-clause *fire* outputting the THEN-clause (i.e., $\psi(x) \Leftarrow \varphi(x)$ and $\varphi(\alpha)$ output $\psi(\alpha)$). The user may once more semantically interpret the output ($\psi(\alpha)$) as an imperative (e.g., $\psi(\alpha)$ *must* be the case), but the algorithm does not *per se* output it in an imperative or any other modal form. It can also be the case that it outputs it as a conditional with empty antecedent ($\psi(\alpha) \Leftarrow$) (e.g., Gebser et al. 2012, §2.2). Note that such *translations* from one modality to another should be performed with high cautiousness. Take for instance the topical logical problem of translating propositional modality to deontic modality whose operators $\mathcal{O}$ & $\mathcal{P}$ are considered to be *hyperintentional*. By "*hyperintentional*", one means that *logically equivalent* conditionals of propositional modality can not be substituted in the scope of deontic operators *salva veritae*.[29] E.g., we can not translate "`shake_hands(pope,Shakira)` $\longleftrightarrow$ `shake_hands(Shakira, pope)`" to "$\mathcal{O}$`shake_hands(pope,Shakira)` $\longleftrightarrow$ $\mathcal{O}$`shake_hands(Shakira, pope)`" (Faroldi 2019, p.388-399; *cf.* Berto and Nolan 2021).

Let's see a toy example of how the application of a specific legal provision to particular facts can be translated into a subsumptive-deductive inference. The language used is that of the *designing* phase of engineering a model ($\mathcal{L}_d$). We can also translate it to a specific programming language $\mathcal{L}_b$ to be used by a specific machine $\mathcal{M}^{\mathcal{L}_b}$ (e.g., my laptop). In Example 2 below, we will see such a translation from an actual example of legal AI that uses a variation of the logic programming PROLOG, probably the most well-known logical programming language (*cf.* Bratko 1990).

Example 1.
According to the ECtHR's case-law (ECtHR Registry 2021, §III.B), whenever an HCP's interference with an individual's freedom of expression is *prescribed by law* ($\mathcal{C}_1$), it *is necessary in a democratic society* ($\mathcal{C}_2$), and it *is in the interests of a legitimate aim* ($\mathcal{C}_3$), then that interference does *not* constitute a violation of ARTICLE 10 (FREEDOM OF EXPRESSION) (ψ). Otherwise, it does, and hence, the HCP has the negative *obligation* to not interfere with the right to freedom of expression. Disobeying that obligation and interfering with the right to freedom of expression constitutes a *violation* of the Convention.

ARTICLE 10
FREEDOM OF EXPRESSION

> *2. The exercise of [freedom of expression], since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.*

For instance, in the *Garaudy v. France (2003)* case, France interfered with the applicant's right to freedom of expression by convicting the applicant for the "*offences of disputing the existence of crimes against humanity, defamation in public of a group of persons – in this case, the Jewish community – and incitement to racial hatred*" (CoE 2018, p.4). The Court judged that this interference was legitimate since it satisfied all three conditions. However, in another case of historical negationism, that of *Perincek v. Switzerland (2015)*, the Court judged that while the conditions $\mathcal{C}_1$ & $\mathcal{C}_3$ were satisfied, the interference was *not* necessary for a democratic society ($\neg\mathcal{C}_2$).

From these case-law examples, we can derive the following imperative: $\psi(x) \Leftarrow \neg\mathcal{C}_1(x) \vee \neg\mathcal{C}_2(x) \vee \neg\mathcal{C}_3(x)$, where x is a variable that can take the form of specific state-interferences (e.g., the Swiss state criminally convicting the applicant in the *Perincek v. Switzerland* case). For the state interference a in the *Perincek v. Switzerland*

[28] Hilpinen and McNamara 2021, §5; McNamara and Putte 2022, §2.
[29] The general definition of *hyperintentional* concepts (e.g., hyperintentional operators) in the SEP is that "*[a] **hyperintensional** concept*

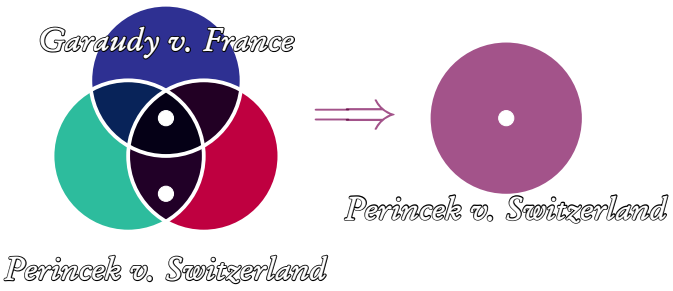


Figure 3: Graphical representation of how *interpretation* can be modelled using *subsumption-deduction* in *interpretive* concepts. The colours represent: • prescribed by law; • legitimate aim; • necessary for a democratic society; • violation of the Convention.

(2015) case, we have that $\mathcal{C}_1(a) \wedge \neg\mathcal{C}_2(a) \wedge \mathcal{C}_3(a)$ and consequently we get the output $\psi(a)$ (the Convention was violated). For the state interference b in the *Garaudy v. France (2003)*, we have that $\mathcal{C}_1(b) \wedge \mathcal{C}_2(b) \wedge \mathcal{C}_3(b)$ and consequently we do not get the output $\psi(b)$. As I will argue in §IV.2.1.1, ¶3, in legal AI, it is common practice to consider propositions which are not made explicitly true as false *by default*. This is premissed on the *principle of the presumption of innocence* according to which the applicant has to *prove* that the defendant is legally responsible for harming them. Consequently, $\psi(b)$ is false by default and hence we have no violation of the Convention.

The subsumptive-deductive inference schema can be used as a model of judicial *justifications* since it reflects how the law is applied to the particular facts of a case (*cf.* §I.3.2.1.2). It is *not* though the only type of judicial justification. As already argued, there is a plethora of other reasoning methods that are used by judicial authorities to justify their judgements.[30] In CHAPTER IV, we will see how a variation of the subsumptive-inference can be used to model *causal* justifications.

Let's see now how Vaquero's five models of legal science can be distinguished based on the foregoing logical construal of interpreting & applying the law (Vaquero 2013, §2). All 5 models can be classified into two categories: those for which the subsumptive tests are *complete* and always *rationally* justified (i.e., grounded on a specific inference method of human reason) and those which are *incomplete* and/or *non*-rational. When it comes to descriptive legal science, the *normativistic* model of legal science is the construal of legal science as the identification of what *is* expected by the legal tradition regarding the interpretation and application of the law in terms of *rational justification*: describing the norms of the legal tradition, systematising its legitimate rational justification methods, identifying logical flaws in judgements, etc. On the contrary, the *realistic* model is about identifying the extraneous *non*-rational factors that influence the *actual* judicial decision-making like the examples of biases in §I.3.1. As one can see by the example of the *hungry judge effect* in §3.1.2.2, the realistic model of legal science belongs to the subject matter of disciplines like experimental psychology, sociology of law, and judicial politics, while the normativist model belongs in the subject matter of disciplines like formal philosophy & logic which identify and formalise *rational* reasoning methods. All those disciplines can contribute to ALGOAI engineering by helping to separate the wheat (rational justifications) from the chaff (non-rational justifications). The normativistic descriptive model though is a *necessary* aspect of legal ALGOAI engineering since it is the model responsible for identifying the reasoning methods required to realise the value of the *rule of law* (§§I.3.2.1.2-3.3).

Which is the appropriate *descriptive* model of legal science?

What about the normative models of legal science? They can be differentiated based on the different approaches they have to the implications of Benacerraf's curse to the interpretation of law: can we decide whether a particular is subsumed by a concept when we lack adequate *prior* inter-subjective knowledge? On the one

draws a distinction between necessarily equivalent contents. If the concept is expressed by an operator, H, then H is hyperintensional insofar as HA and HB can differ in truth value in spite of A and B's being necessarily equivalent." (Berto and Nolan 2021).

[30] In what follows, I provide citations on different types of judicial reasoning that I consulted for this Thesis. Their classification is not strict since many citations overlap with multiple types of judicial reaosning. Overviews of different judicial reasoning types: Eisenberg 2022; Bongiovanni et al. 2018; Gold 2018; Armgardt, Canivez, and Chassagnard-Pinet 2015; Hage 2005; Horovitz 1972; Walton 2002; Causal inference: Moore 2009, 2019; Stoyanova 2018; Schaffer 2000; Lavrysen 2018; Shafer 2002; Hart and Honoré 1959, 1985; Wright 1985, 1988, 2011; Turton 2020; Green 2015; Plakokefalos 2015; Sulyok 2017; Non-monotonic reasoning: Sartor 2012; Rigoni 2014; Poggi 2021; Gordon 1988; Analogical reaosning: Lamond 2016; Emmert 1992; Deontic logic: Navarro and Rodríguez 2014; Governatori, Rotolo, and Sartor 2021; Sven Oven Hansson 2021; Royakkers 1998; Canavotto 2020; van Woerkom et al. 2022; von Wright 1951; Hilpinen and McNamara 2021; McNamara and Putte 2022; More logic: Alchourrón 2015; Prakken 1993; Haack 2007; Legal interpretation: Greenberg 2021; Neves 2021; Dworkin 1986, 2011; Stavropoulos 1996; Schroeter, Schroeter, and Toh 2020; Iatrou 2022a; Letsas 2013; von der Lieth Gardner 1987; Schauer 1991; Solan and Tiersma 2012, Part II; Others: Alexander and Sherwin 2008; MacCormick 1992; Sartor 2009; Koziol 2015.

hand, there are those legal scientists that attempt to craft normative methods of rational justification that can ideally respond to every challenge induced by Benacerraf's curse: we can always rationally justify the interpretation and subsequent application of the law (*argumentativist* model). Such an example is the method of *reflective equilibrium* that we will see in §§III.3.2.1,III.3.3. On the other hand, there are those that reject this rational "*absolutism*". Such a model is the *realist-technological* normative model of legal science whose main premiss is essentially conceding that in certain cases, Benacerraf's curse is unsolvable, and hence, we need to resolve to non-rational means of deciding the interpretation & application of a concept. Such an example is the case of *decisionism* that we saw in §I.3.2.1.2, ¶7. Finally, for the *critical* model, the legal scientists should go *political*: "*the law is a continuation of politics by other means, thus legal scholars are political agents that must be aware of the important role they play, and they should act accordingly.*" (¶61; *cf.* Cahoone 2023). Take for instance the Venice Commission's remark on the rule of law in the eastern European countries: "*[t]he notion of the rule of law is however often difficult to find in former socialist countries which experienced the notion of* ***socialist legality****.*" (CDL-AD(2011)003rev, ¶33). Or the conflict between judicial review and the majority rule in illiberal democracies that we saw in §I.2.8, ¶2.

Which is the appropriate *normative* model of legal science?

In the case of the ECtHR, we should use a combination of the *argumentativist* & the *critical* models. The justification for the argumentativist model is pretty straightforward: the ideal for the rule of law is to identify *rational* justifications for *every* case of interpretation (§§I.3.2.1.2-3.3). E.g., this way we can predict in an inter-subjective way the application of the law corroborating the legitimacy value of foreseeability. Having said that, we still have to go *political*. For the ECtHR, it must be the case that its judgements are on par with *present-day* standards shared by the High Contracting Parties (Letsas 2013, pp.108-109), in contrast with other interpretation approaches like *originalism* according to which interpretation should emphasise aspects of the Convention at the time that it was ratified (e.g., intentions of the framers or ratifiers and/or how legal interpretation was performed at the time) (Greenberg 2021, §3). This is why the ECtHR's method of interpreting the Convention is called the *living instrument doctrine*: the Convention is an *instrument* that is *alive* and continuously *evolving*.[31,32] The changes of the "*the present-day standards*" that the Court should take into consideration are the result of certain *political*, *ethical* and *social* theories and practices. For instance, new family models (e.g., patchwork families, one-parent-families), advancements in LGBTQI+ rights, prioritisation of environmental sustainability & data protection (Nussberger 2020, pp.77-82; *cf.* §I.2.6.1, *fn.* 68). Hence, the consideration of political, ethical and social theories and practices is a requisite for the interpretation of the Convention.[33]

interpreting the Convention: going *political*

Now that I introduced what are both legal sciences & AI engineering, I can finally introduce how logic does (not) glue them in the legal AI engineering practice.

[31]For a thorough introduction to the *living instrument doctrine* have a look at Letsas 2013, §3; Nussberger 2020, §3. For a concise summary accessible to the non-expert audience have a look at ECHR's Public Relations Unit 2022. For a comparison among the living instrument doctrine and other methods of interpretation have a look at Letsas 2007, §3.

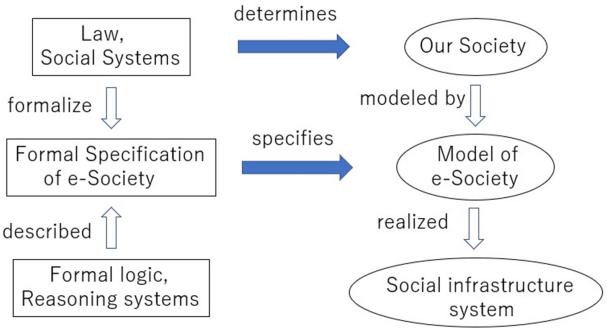
[32]The Court's first *explicitly* acknowledgement of the living instrument doctrine happened in the *Tyrer v. United Kingdom (1978)* case: "*The Court must also recall that the Convention is a* ***living instrument*** *which, as the Commission rightly stressed, must be interpreted in the light of* ***present-day*** *conditions. In the case now before it the Court cannot but be influenced by the developments and* ***commonly accepted standards*** *in the penal policy of the member States of the Council of Europe in this field.*" (¶31, emphasis added). In this case, the Court had to decide whether corporal punishment of juveniles (bare-skin birching) that was prescribed by law in the Isle of Man (dependent territory of the UK) violates ARTICLE 3 (PROHIBITION OF TORTURE). The Court took into consideration for its final judgement (violation of ARTICLE 3) that the vast majority of the HCPs at the time including the rest of the UK had abolished corporal punishment (Letsas 2013, p.109-110). Note, that contra to Letsas, Christoffersen and Madsen 2011 (§3) argue that the living instrument doctrine was not used from the Court's early days and that this novel change in the 70s took the High Contracting Parties "*by surprise*" and signaled "*a new beginning for European human rights*" (*ibid.*, p.7).

[33]*Cf.* Stavropoulos 1996, pp.50-51; Dworkin 2011, p.120.

II.4.2 Gluing the prickles of hedgehogs: logic & legal AI

In this section, I will show how logic *does* (not) glue the two hedgehog-disciplines and it *should* glue them so as to engineer *legitimate* legal ALGOAI. I begin by introducing logic-based legal AI, I continue with connectionist legal AI, and I conclude with hybrid logic-based & connectionist models, the future of legal ALGOAI.

II.4.2.1 Logic-based legal AI



Japan's e-Society project (2004-2010): a new social order ordered by formalised laws. (*source:* Nitta and Satoh 2020, Figure 8)

The idea that we can engineer a new social order by using formal reasoning to model the laws of that order is already present in Japan's *e-Society* project facilitated by Japan's Ministry of Education, Culture, Sports, Science and Technology and the Japan Advanced Institute of Science and Technology (2004-2010). "*At the preliminary stage of the project, the project analyzed various laws, regulations, and social customs as logical models and it was thought that the e-Society model would be regarded as one of the specifications of a high-level information social system*" (Nitta and Satoh 2020, p.485; *see* Figure on the left). Nitta and Satoh attribute e-Society's failure to the project being too big for a single research team to handle (*ibid.*, p.486). Albeit this being true, I would argue that it was also too early technological-wise since the elements that make SOCIETY 5.0 an ambitious but still achievable prospect are the fruits of INDUSTRY 4.0 which came to happen a few years after the end of the e-Society project (§I.2.6.1).

Probably the two most characteristic types of logic-based legal AI are *rule-based reasoning (RBR)* and *case-based reasoning (CBR)* expert systems (*cf.* Prakken 1993, §2.4). RBR is a direct logical modelling of the *subsumptive-deductive* model of judicial reasoning using IF-THEN rules (§4.1.2). In particular, an RBR model is a conjunction of IF-THEN rules that represent laws. For each case, we provide as *input* to the model certain facts, and then, the rules that have those facts as their antecedent *fire* outputting their consequents. The formal language $\mathcal{L}_b$ used to build such models is usually that of first-order logic since rules need to be applicable in many particular cases and hence the existence of *universals* (i.e., *predicates*) in their IF-clauses becomes a necessity (*cf.* MacCormick 1992, p.186). Look for instance the following real-life example of a legal expert system's architecture. The expert system is KRIP, a Japanese legal AI model of consultation for patent law which was developed in the 80's (Yoshino 1987) and that was built using a PROLOG-based first-order logical language $\mathcal{L}_b$:

rule-based reasoning (RBR)

Example 2.

LEGAL PROVISION

"***Section 7 (1)*** *Minors or adult wards may not undertake procedures except through their statutory representatives;*"

PROLOG-based logical language $\mathcal{L}_b$

```
procedural_ability_person(X) :- not minor(X), not adult_ward(X),!.
procedural_ability_person(X) :- statutory_representative(X).
```

In this example, Section 7(1) of the patent law is translated to two IF-THEN rules the subsumptive-deductive inference schema of §4.1.2. The ":-" symbol is used in the place of "⇐", the predicate `minor/1` represents the property (concept) of *being a minor*, the predicate `statutory_representative/1` represents the property (concept) of *being a statutory representative*, and so forth.

For an overview of the use of RBR systems in Dutch legal practice *see* Timmer and Rietveld 2019. For contemporary examples of legal RBR using Answer Set Progamming (ASP) non-monotonic first-order logical languages $\mathcal{L}_b$ *see* Morris 2021; Cabalar, Fandinno, and Fink 2014; Wan, Kifer, and Grosof 2015; Aravanis, Demiris, and Peppas 2018; Iatrou 2022b, 2022a; *cf.* Lifschitz 2019; Gebser et al. 2012. A landmark case of RBR is the use of PROLOG to model the British Nationality Act (Sergot et al. 1986). For an interesting overview of the challenges RBR legal expert systems faced during their peak in the 80's that hunt us until today *see* von der Lieth Gardner 1987; Linant de Bellefonds 1994, as well as, the proceedings of the "*Expert systems in law:*

Impacts on legal theory and computer law" workshop that took place in Tübingen, Germany in 1988 (Fiedler, Haft, and Traunmüller 1988). For literature on the theoretical logical foundations of RBR (and expert systems) *see* MacCormick 1992; Rigoni 2014; Alchourrón 2015.

What about CBR? As already argued (§I.2.1, *fn.*24; §I.2.4, ¶2), CBR is a type of *analogical reasoning* which is based on the principle of equality: similar cases should be treated similarly and dissimilar cases should be treated dissimilarity. A traditional approach to logic-based CBR is *defeasible reasoning* like the examples we saw in §I.2.4, ¶3. Take for instance the model of Liu et al. 2022. In that model, there are two opposing sides, the plaintiff & the defendant, and they are on a trial about whether a specific law has been violated. The plaintiff uses as arguments a set of atoms $\mathcal{A}_1$ that represent facts and the defendant another set of atoms $\mathcal{A}_2$. At the same time, there is a knowledge base CB of arguments used in past cases and a *preference relation* $\prec_{CB}$ among those arguments based on what judicial authorities have decided about those past cases. That preference relation can be used to judge which set of the current case's arguments ($\mathcal{A}_1$ and $\mathcal{A}_2$) will *defeat* the other and hence why the types of logic used in such models are called *defeasible logics* (for a similar CBR approach *see* van Woerkom et al. 2022). Two landmark legal AI models, HYPO (1984; *see* Ashley 1990) & CATO (1997; *see* Aleven 1997), are paradigmatic examples of CBR (*cf.* Roth 2003; Nitta and Satoh 2020, figure 1). For a concise introduction to CBR that includes analysis and comparison between HYPO and CATO *see* Roth 2003. The reader may also be interested in John F. Horty's contributions to the CBR literature which constitutes a reference point for contemporary CBR (Horty and Bench-Capon 2012; Horty 2011, 2004).

case-based reasoning (CBR)

Finally, there are always hybrid cases like the HELIC-II model (Yoshino 1998) that combines RBR and CBR. The model first uses RBR to apply the law and in case of inconsistencies in the law application it uses CBR to see how those inconsistencies were resolved in the past (Nitta and Satoh 2020, p.479).

II.4.2.2 Connectionist legal AI

We saw in §I.3.2.1.1, *fn.*81 that NN-based AI is designed by being trained to identify patterns in large amounts of data (training data) so as to use those patterns to produce new information. A widely cited NN-based legal AI model, and to my knowledge the most cited when it comes to ECtHR predictive justice, is Aletras et al.'s 2016 *natural language processing* (*NLP*) model[34] which is trained by identifying patterns in past ECtHR judgements about violations of the Convention.[35] and then it uses those patterns to classify new cases as (non-)violations of the Convention. In 2019, more or less the same team of experts trained more connectionist AI models using the so-called *attention mechanisms* with many of those new models achieving higher performance than the 2016 model (for more literature on connectionist AI about ECtHR case-law *see*: Moreira 2022; Medvedeva et al. 2020; Kaur and Božzić 2020). Note that CHATGPT and many of the groundbreaking generative AI models do use a specific type of attention mechanism with its foundational paper being Vaswani et al.'s 2017 "*Attention is all you need*".

Attention is another AI method that is inspired by human cognitive abilities (*cf.* §I.3.2.1.1, ¶10), the cognitive ability to draw attention only to those aspects of new information which are of relevance to the *intended application* of that information while disregarding the rest (*cf.* §III.1, ¶4). In the example of Chalkidis, Androutsopoulos, and Aletras 2019, the following visual representation of how attention works called *heatmap* sheds light on how this human-based mechanism operates:

[34]NLP is "*is the [subdiscipline] of computer science which studies how to equip computer systems to handle the language naturally spoken by humans. NLP techniques leverage linguistic studies in natural language and address the syntax, semantics, and pragmatics (the contextual meaning) of expressions in natural language.*" (Sartor and Loreggia 2020, p.41). An application of NLP that shows the positive impact that ALGOAI can have the legitimacy of SOCIETY 5.0's cyber-physical political order is *upload filters* that decide which social acts can be performed in the *cyber* component of the cyber-physical order (§I.2.6.1, *fn.*70).

[35]The model is fed with the ECtHR's documents of past judgements which contain the facts of a case and the Court's interpretation & application of the Convention on those facts. The documents of past judgements are found in the ECtHR's public database HUDOC.

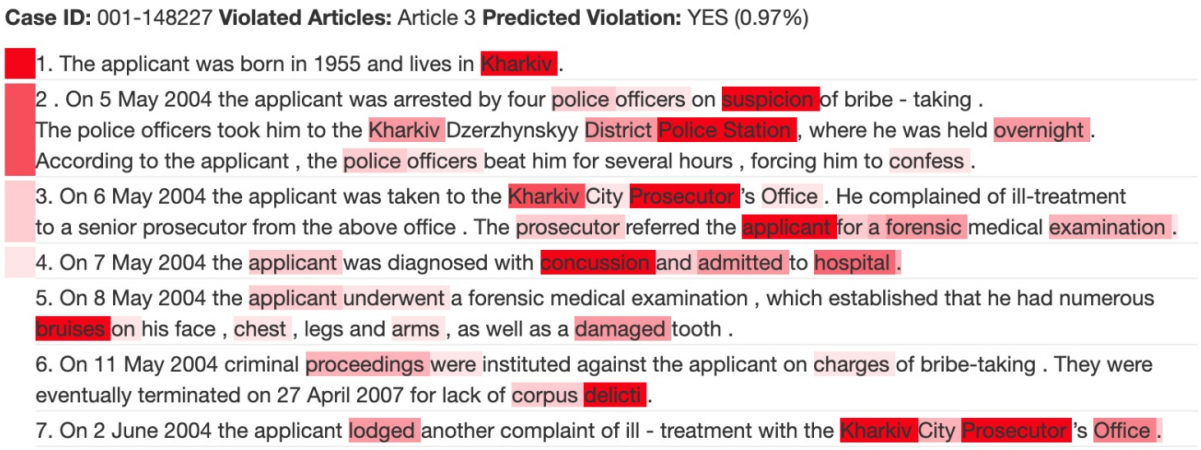


Figure 4: Attention over words (colored words) and facts (vertical heat bars) as produced by HAN. The "*reder*" the color, the more the attention!! (Chalkidis, Androutsopoulos, and Aletras 2019 Figure 1).

More precisely, this is an example of a heatmap for a HIERARCHICAL ATTENTION NETWORK (HAN)[36] trained by Chalkidis, Androutsopoulos, and Aletras Chalkidis, Androutsopoulos, and Aletras that performed quite well in discerning Convention violations with non-violations. The model uses two types of attention. The first type is attention drawn to specific *words* in the facts of a case that raise the probability of that case being a violation of the Convention. They are the words highlighted in red in Figure 4: the more faded the red the less attention is drawn. The second type of attention is the one drawn to individual *facts* as a whole and not just to words of those facts. In Figure 4, each fact is numbered and the intensity of red on the left of the respective number shows how much attention is drawn to that particular fact. The higher the attention, the higher the contribution of that particular fact in classifying the case it describes as a violation of the Convention.

Albeit attention heatmaps are a popular technique to "*open*" the black-box of connectionist AI and explain its output (*cf.* §I.3.2), one could hardly classify the as *meaningful* explanations. As the reader will see by themselves, the highlighted words & facts can hardly justify why the law was applied the way it was. For instance, the fact that the model identified a strong *statistical correlation* between the word "*bruises*" and the the violation of an article about torture says absolutely nothing about the application of the law. It may be common that in cases of violation of ARTICLE 3 the world bruises appeared frequently, but the mere appearance of bruises can not justify an application of the law. Who caused the bruises? Under which circumstances? And so on (*cf.* ECtHR's Registry 2022). As Ulenaers 2020 argued wittily (p.27):

the non-sense of connctionist AI

> "*if a defendant wishes to know the reasons why they were convicted, they have the right to get a better answer than "we trained the system on lots of data, and this is what it decided" (Tegmark 2018, p.106)*"

Even worse, since statistical correlation are predicated on the *frequency* of patterns that appear in the training set, connectionist AI models are susceptible to incorporate biases that appear in past judgement and to reinforce the present-day dominant views of how the law should be applied stagnating progress. (*cf.* Gordon et al. 2022; §I.2.5, ¶8). Take for instance the example of the word "*Kharkiv*" in Figure 4. The fact that this word appeared frequently in many past violations should not make the Court *biased* towards applications lodged against Ukraine. Each case should be judged independently. Hence why Chalkidis, Androutsopoulos, and Aletras 2019 retrained their model by blanking this word out.

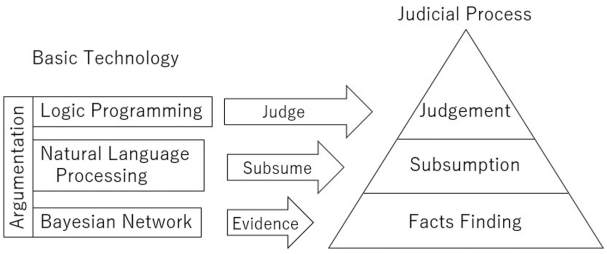
In general, the wide success of connectionist AI (*see e.g.* §I.3.1) and its subsequent extended use has already raised ethical, legal, political, and practical concerns about its explainability spawning a new subdiscipline of the AI discipline, that of *explainable AI* (*XAI*). XAI's subject matter is AI architectures that provide *justifications* for the AI's output (Angelov et al. 2021; Gunning et al. 2019). The emergence of the XAI subdiscipline is another example of LAWS 3.0 & 4.0 attempting to resolve *legitimacy* concerns induced by new ALGO technologies (*cf.* Commissioner for Human Rights 2019; CEPEJ 2019; Goodman and Flaxman 2017; MSI-AUT 2019). Concerns that are shared by legal practitioners themselves who subsequently advocate for further opening of the black-boxes (*see e.g.* Górski and Ramakrishna 2021; Adrien et al. 2021).

XAI & legitimacy

[36] It is state-of-the-art (at least until 2019) AI model used in text classification that was introduced by Yang et al. 2016 (Chalkidis, Androutsopoulos, and Aletras 2019).

II.4.2.3 Hybrid legal AI: the future (?)

Considering the above, it is the opinion of the author that legal ALGOAI engineers should try to combine the different AI architectures in an effort to maximize their advantages and balance out their limitations, what is called in the literature *hybrid AI*. The basis of hybrid ALGOAI's architecture should be a *connectionist* architecture in order to be able regulate *massive* amount of data in *reasonable* time, reasonable enough to satisfy the legitimacy requirement of legality in the context of the cyber-physical SOCIETY 5.0. Some components of that connectionist architecture though should be modified so as to accommodate logic-based architecture that enforces specific *logical* structures to the *justifications* provided by ALGOAI for its output. One could counterargue that we do not need a logic-based AI architecture to generate such explanations. Take for instance the example of the Colombian judge Juan Manuel Padilla using CHATGPT as supportive AI to write a judgement about Colombian insurance law enthusiastically approving of its judgement (Taylor 2023 *contra* argument in §I.3.2.1, ¶1 about the bias of accepting AI judgements that would not have been accepted had the AI not be used). Indeed, *prima facie*, as long as they *look* like legitimate justifications, they can be used by judicial authorities. However, their probabilistic character always leaves room for mistakes, even if they do occur rarely ones (*cf.* with the comparison of accuracy between connectionist AI and logic-based AI that solve SUDOKU puzzles in §I.3.2.1.1, ¶12). Rigid, *a priori* logical structures force into ALGOAI's justification a specific logical structure, the logical structure the legitimate authorities have decided in advance. There is no room for misorientation.



Advanced Reasoning Support for Judicial Judgment by Artificial Intelligence project (2017-Present): different AI architectures for different tasks in the process of interpreting & applying the law (*source:* Nitta and Satoh 2020, Figure 9)

The gist of my proposal is to segment different parts of judicial reaosning and allocate each part to a different type of AI architecture like the example in the Figure on the left, another project supported by the Japanese government under the supervision of prof. Ken Satoh of Japan's National Institute of Informatics. As one can see in that example, *logical* structures are embedded in the *last* part of the input's process, after using connectionist AI to determine the facts of the cases and which concepts subsume those facts. This way, we make sure that the final judgement will have a structure that reflects specific types of judicial reasoning. The method of forcing a specific logical structure at the end of a process performed by connectionist AI is a hybrid AI method called *pipeline* (Giunchiglia, Stoian, and Lukasiewicz 2022, p.5481; for an example of a pipeline used to solve SUDOKU *see* Yang, Ishay, and Lee 2020). Note that such logical constraints can be placed in any part of connectionist AI. They can also be placed in the input (e.g., forcing specific symbolic structures on the facts of a case) or in the AI's components that *process* the input (e.g., forcing specific anthropomorphic symbolic ways to combine the facts from the input with laws that are entrenched in the model) (Giunchiglia, Stoian, and Lukasiewicz 2022). An interesting approach on using logic-based CBR in order to impose normative requirements in black-box AI models like COMPAS is van Woerkom et al. 2022. Other examples with the use of formal structures to open black-boxes are: Sivaram 2022; Baron 2023; Francis Rhys Ward and Belardinelli 2022; Beckers 2022; O'Shaughnessy et al. 2020; Neelakantan 2017; Towell and Shavlik 1994.

II.5 Up for the META!

Summing up, the process of ALGOAI engineering should be segmented into three phases: the *designing*, the *building*, and the *evaluation* of the model. Part of that process is the identification of judicial reasoning methods and the subsequent *translation* to formal components of AI models. In the process of translating, the ALGOAI engineers will have to *cross* the borders of their disciplines and collaborate with experts from disciplines with substantial differences in their *interpretation* of the world; different semantics, different theories, different methodologies, and more importantly different purposes. It is the role of the logician & the formal philosopher to make sense of this babel and glue them into one coherent *legitimate meta*-disciplinary practice. In the next CHAPTER, we will see an example of a methodology that can be used by ALGOAI engineers to produce formal translations of judicial reaosning as well as how the logician & the formal philosophers can fulfill their role as *foxes*.

References

Adrien, Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. 2021. "Legal requirements on explainability in machine learning." *Artificial Intelligence and Law* 29:149–169. https://doi.org/10.1007/s10506-020-09270-4.

Alchourrón, Carlos E. 2015. "Limits of logic and legal reasoning." In *Essays in legal philosophy*, Reprint, edited by Carlos Bernal and Carla Huerta. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198729365.003.0017.

Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. 2016. "Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective." *PeerJ Computer Science* 2:e93.

Aleven, Vincent. 1997. "Teaching case-based reasoning through a model and examples." PhD diss., University of Pittsburgh.

Alexander, Larry, and Emily Sherwin. 2008. *Demystifying legal reasoning.* Cambridge Introductions to Philosophy and Law series. Cambridge University Press.

Angelov, Plamen P., Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. 2021. "Explainable artificial intelligence: An analytical review." *WIREs Data Mining and Knowledge Discovery* 11 (5): e1424. https://doi.org/10.1002/widm.1424.

Angius, Nicola, Giuseppe Primiero, and Raymond Turner. 2021. "The philosophy of computer science." In *The Stanford Encyclopedia of Philosophy*, Spring 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Apostel, Leo, Guy Berger, Asa Briggs, and Guy Michaud, eds. 1972. In *Interdisciplinarity: Problems of teaching and research in universities.* OECD Publications Center.

Aravanis, Theofanis, Konstantinos Demiris, and Pavlos Peppas. 2018. "Legal reasoning in answer set programming." *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, 302–306.

Armgardt, Matthias, Patrice Canivez, and Sandrine Chassagnard-Pinet, eds. 2015. *Past and present interactions in legal reasoning and logic.* Vol. 7. Logic, Argumentation and Reasoning. Springer Cham. https://doi.org/10.1007/978-3-319-16021-4.

Arnold, Markus. 2020. "Interdisciplinary research (Interdisciplinarity)." In *Encyclopedia of creativity, invention, innovation and entrepreneurship*, 2nd ed., edited by Elias G. Carayannis. Springer Cham.

Ashley, Kevin D. 1990. *Modeling legal argument: Reasoning with cases and hypotheticals.* MIT Press.

Baron, Sam. 2023. "Explainable AI and causal understanding: Counterfactual approaches considered." *Minds and Machines* 33 (2): 347–377. https://doi.org/10.1007/s11023-023-09637-x.

Beckers, Sander. 2022. "Causal explanations and XAI." *Proceedings of Machine Learning Research 1st Conference on Causal Learning and Reasoning* 140:1–20.

Benacerraf, Paul. 1983. "Mathematical truth." In *Philosophy of mathematics: Selected readings*, 2nd ed., edited by Paul Benacerraf and Hilary Putnam. Cambridge University Press.

Berlin, Isaiah. (1953) 1954. *The hedgehog and the fox: An essay on Tolstoy's view of history.* Reprint. Weidenfeld and Nicolson.

Berlin, Isaiah, and Ramin Jahanbegloo. 1991. *Conversations with Isaiah Berlin.* Charles Scribner's Sons.

Berto, Francesco, and Daniel Nolan. 2021. "Hyperintensionality." In *The Stanford Encyclopedia of Philosophy*, Summer 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Boden, Margaret A., ed. 1990. *The philosophy of artificial intelligence.* Oxford readings in philosophy. Oxford University Press.

Bongiovanni, Giorgio, Gerald Postema, Antonino Rotolo, Giovanni Sartor, Chiara Valentini, and Douglas Walton, eds. 2018. *Handbook of legal reasoning and argumentation.* Springer, Dordrecht.

Bratko, Ivan. 1990. *PROLOG: Programming for Artificial Intelligence.* 2nd ed. Edited by Andrew D. McGettrick and Jan van Leeuwan. International Computer Science. Addison-Wesley Publishing Company.

Britannica, The Editors of Encyclopaedia. 2023. *Sir Isaiah Berlin.* Online ed. Encyclopedia Britannica, March 17, 2023. Accessed March 24, 2023. https://www.britannica.com/biography/Isaiah-Berlin.

Cabalar, Pedro, Jorge Fandinno, and Michael Fink. 2014. "Causal graph justifications of logic programs." *Theory and Practice of Logic Programming* 14 (4-5): 603–618. https://doi.org/10.1017/S1471068414000234.

Cahoone, Lawrence. 2023. "The end of Enlightenment liberalism?" *The Journal of Speculative Philosophy* 37 (1): 81–98.

Campbell, David F. J., and Wieland Schneider. 2020. "Media and innovation." In *Encyclopedia of creativity, invention, innovation and entrepreneurship,* 2nd ed., edited by Elias G. Carayannis. Springer Cham.

Canavotto, Ilaria. 2020. "Where responsibility takes you: Logics of agency, counterfactuals and norms." PhD diss., Institute of Logic, Language and Computation, Universiteit van Amsterdam.

Capaldi, Nicholas. 1998. *The Enlightenment project in the analytic conversation.* Edited by H. Tristram Engelhardt. Vol. 4. Philosophical Studies in Contemporary Culture. Springer Netherlands. https://doi.org/10.1007/978-94-017-3300-7_1.

CEPEJ (European Commission for the Efficiency of Justice). 2019. *European ethical Charter on the use of artificial intelligence in judicial systems and their environment.* Printed by the Council of Europe. https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment.

Chalkidis, Ilias, Ion Androutsopoulos, and Nikolaos Aletras. 2019. "Neural legal judgment prediction in English." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 4317–4323. Florence, Italy: Association for Computational Linguistics, July. https://doi.org/10.18653/v1/P19-1424.

Chalmers, David J. 2020. "What is conceptual engineering and what should it be?" *Inquiry: an interdisciplinary journal of philosophy,* 1–18. https://doi.org/10.1080/0020174X.2020.1817141.

Chatziathanasiou, Konstantin. 2022. "Beware the lure of narratives: "hungry judges" should not motivate the use of "artificial intelligence" in law." *German Law Journal* 23 (4): 452–464. https://doi.org/10.1017/glj.2022.32.

Cherniss, Joshua, and Henry Hardy. 2022. "Isaiah Berlin." In *The Stanford Encyclopedia of Philosophy,* Summer 2022, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Christoffersen, Jonas, and Mikael Rask Madsen, eds. 2011. *The European Court of Human Rights: Between law and politics.* Online 1st ed. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199694495.001.0001.

CoE (Council of Europe). 2018. *'Memory laws' and freedom of expression.* Thematic Factsheet. Platform to promote the protection of journalism / safety of journalists, July. Accessed January 21, 2023.

Commissioner for Human Rights. 2019. *Unboxing artificial intelligence: 10 steps to protect Human Rights.* Printed at the Council of Europe, May.

Crow, Graham. 2020. "Hedgehogs, foxes and other embodiments of academics' research career trajectories." *Contemporary Social Science* 15 (5): 577–594. https://doi.org/10.1080/21582041.2020.1849784.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous factors in judicial decisions." *Proceedings of the National Academy of Sciences* 108 (17): 6889–6892. https://doi.org/10.1073/pnas.1018033108.

de Jong, Willem R., and Arianna Betti. 2010. "The classical model of science: a millennia-old model of scientific rationality." *Synthese* 174 (2): 185–203. https://doi.org/10.1007/s11229-008-9417-4.

Dworkin, Ronald. 1986. *Law's empire.* Harvard University Press.

———. 2011. *Justice for hedgehogs.* Belknap Press.

ECHR's Public Relations Unit. 2022. *The European Convention of Human Rights: A living instrument.* https://www.echr.coe.int/Documents/Convention_Instrument_ENG.pdf.

ECtHR Registry. 2021. *Guide on Article 10 of the European Convention on Human Rights: Freedom of expression.* Updated. April. https://www.echr.coe.int/documents/guide_art_10_eng.pdf.

ECtHR's Registry. 2022. *Guide on Article 3 of the European Convention on Human Rights: Prohibition of torture.* Updated. August.

Eisenberg, Melvin A. 2022. *Legal reasoning.* Cambridge University Press. https://doi.org/10.1017/9781009162517.

Emmert, Craig F. 1992. "An integrated case-related model of judicial decision making: Explaining state Supreme Court decisions in judicial review cases." *The Journal of Politics* 54 (2): 543–552.

Faroldi, Federico L. G. 2019. "Deontic modals and hyperintensionality." *Logic Journal of the IGPL* 27 (4): 387–410. https://doi.org/10.1093/jigpal/jzz011.

Fiedler, Herbert, Fritjof Haft, and Roland Traunmüller, eds. 1988. *Expert systems in law: Impacts on legal theory and computer law.* Vol. 4. Neue methoden im Recht. Attempto Verlag Tübingen GmbH.

Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli. 2022. "A causal perspective on AI deception." In *Proceedings of the International Conference on Logic Programming 2022 (ICLP 2022)*, edited by Joaquín Arias, Roberta Calegari, Luke Dickens, Wolfgang Faber, Jorge Fandinno, Gopal Gupta, Markus Hecher, et al. Haifa, Israel, July.

Frisch, Mathias. 2022. "Causation in Physics." In *The Stanford Encyclopedia of Philosophy*, Spring 2022, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Gabbrielli, Maurizio, and Simone Martini. 2010. *Programming Languages: Principles and paradigms.* Translated. *orig: Linguaggi di programmazione: Principi e paradigmi*. Edited by Ian Mackie. Undergraduate Topics in Computer Science (UTiCS). Springer.

Gebser, Martin, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. 2012. "Answer set solving in practice," https://doi.org/10.2200/S00457ED1V01Y201211AIM019.

Giunchiglia, Eleonora, Mihaela Catalina Stoian, and Thomas Lukasiewicz. 2022. "Deep learning with logical constraints." In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, edited by Lud De Raedt, 5478–5485. International Joint Conferences on Artificial Intelligence Organization, July. https://doi.org/10.24963/ijcai.2022/767.

Gold, Michael Evan. 2018. *A primer on legal reasoning.* ILR Press.

Goodman, Biyce, and Seth Flaxman. 2017. "European Union regulations on algorithmic decision-making and a "right to explanation"." *AI Magazine* 38 (3).

Gordon, Mitchell L., Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. "Jury learning: Integrating dissenting voices into machine learning models." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* CHI '22. New Orleans, LA, USA: Association for Computing Machinery. https://doi.org/10.1145/3491102.3502004.

Gordon, Thomas F. 1988. "The importance of nonmonotonicity for legal reasoning." In *Expert systems in law: Impacts on legal theory and computer law*, edited by Herbert Fiedler, Fritjof Haft, and Roland Traunmüller, 4:111–126. Neue methoden im Recht. Attempto Verlag Tübingen GmbH.

Górski, Łukasz, and Shashishekar Ramakrishna. 2021. "Explainable artificial intelligence, lawyer's perspective." In *Proceedings of the eighteenth international conference on artificial intelligence and law*, 60–68. Association for computing machinery. https://doi.org/10.1145/3462757.3466145.

Gould, Stephen Jay. 2003. *The hedgehog, the fox, and the magister's pox: Mending the gap between science and the humanities.* Harvard University Press. https://doi.org/10.4159/harvard.9780674063402.

Governatori, Guido, Antonino Rotolo, and Giovanni Sartor. 2021. "Logic and the law: philosophical foundations, deontics, and defeasible reasoning." Chap. 9 in *Handbook of deontic logic and normative systems*, edited by Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, vol. 2. College Publications.

Green, Sarah. 2015. *Causation in negligence.* Hart Studies in Private Law. Hart Publishing.

Greenberg, Mark. 2021. "Legal interpretation." In *The Stanford Encyclopedia of Philosophy*, Fall 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Gunning, David, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. "XAI—Explainable artificial intelligence." *Science Robotics* 4 (37): eaay7120. https://doi.org/10.1126/scirobotics.aay7120.

Haack, Susan. 2007. "On logic in the law: "something, but not all"." *Ratio Juris* 20 (1): 1–31. https://doi.org/10.1111/j.1467-9337.2007.00330.x.

Hage, Jaap. 2005. *Studies in Legal Logic.* Vol. 70. Law and Philosophy Library. Springer Dordrecht. https://doi.org/10.1007/1-4020-3552-7.

Hansson, Sven Ove. 2021. "Science and pseudo-Science." In *The Stanford Encyclopedia of Philosophy*, Fall 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Hansson, Sven Oven. 2021. "Varieties of permission." In *Handbook of deontic logic and normative systems*, edited by Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, vol. 1. College Publications.

Harpham, Geoffrey Galt. 2015. "Defending disciplines in an interdisciplinary age." *College Literature* 42 (2): 221–240. https://doi.org/10.1353/lit.2015.0018.

Hart, H. L. A. 1961. *The concept of law.* Clarendon Law Series. Oxford University Press.

Hart, H. L. A., and Tony Honoré. 1959. *Causation in the law.* 1st ed. Oxford University Press.

———. 1985. *Causation in the law.* 2nd ed. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198254744.001.0001.

Harvey J., Graff. 2015. *Undisciplining knowledge: Interdisciplinarity in the twentieth century.* John Hopkins University Press.

Heckhausen, Heinz. 1972. "Discipline and interdisciplinarity." In *Interdisciplinarity: Problems of teaching and research in universities*, edited by Leo Apostel, Guy Berger, Asa Briggs, and Guy Michaud, Section 1, 83–88. OECD Publications Center.

Hilpinen, Risto, and Paul McNamara. 2021. "Deontic logic: A historical survey and introduction." In *Handbook of deontic logic and normative systems*, edited by Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, vol. 1, Part I: Background, 3–136. College Publications.

Horovitz, Joseph. 1972. *Law and logic: A critical account of legal argument.* Library of Exact Philosophy (LEP) series. Springer Vienna. https://doi.org/10.1007/978-3-7091-7111-0.

Horty, John F. 2004. "The result model of precedent." *Legal Theory* 10 (1): 19–31.

———. 2011. "Rules and reasons in the theory of precedent." *Legal Theory* 17 (1): 1–33. https://doi.org/10.1017/S1352325211000036.

Horty, John F., and Trevor J. M. Bench-Capon. 2012. "A factor-based definition of precedential constraint." *Artificial intelligence and Law* 20:181–214.

Howell, Kerry E. 2013. *An introduction to the philosophy of methodology.* Sage Publications.

Iatrou, Evan. 2022a. "A normative model of explanation for binary classification legal AI and its implementation on causal explanations of Answer Set Programming." In *Proceedings of the International Conference on Logic Programming 2022 (ICLP 2022)*, edited by Joaquín Arias, Roberta Calegari, Luke Dickens, Wolfgang Faber, Jorge Fandinno, Gopal Gupta, Markus Hecher, et al. Haifa, Israel, July.

———. 2022b. "Non-monotonic rule-based logical programming for modelling legal reasoning: the example of Answer Set Programming." In *Proceedings of the 13th Panhellenic Logic Symposium*, 2:135–144. Volos, Greece, July.

Isaac, Manuel Gustavo. 2020. "How to conceptually engineer conceptual engineering?" *Inquiry: An Interdisciplinary Journal of Philosophy*, 1–24. https://doi.org/10.1080/0020174X.2020.1719881.

Jamie, Cohen-Cole. 2014. *The open mind: Cold war politics and sciences of human nature.* University of Chicago Press.

Jantsch, Erich. (1970) 1972. "Inter- and transdisciplinary university: A systems approach to education and innovation." *Higher Education* 1 (1): 7–37. https://doi.org/10.1007/BF01956879.

Jantsch, Erich. 1972. "Towards interdisciplinarity and transdisciplinarity in education and innovation." In *Interdisciplinarity: Problems of teaching and research in universities*, edited by Leo Apostel, Guy Berger, Asa Briggs, and Guy Michaud, Section 3, 97–120. OECD Publications Center.

Johnson, Scott. 1996. "Husbands and wives: A correction." *Contemporary Family Therapy* 18 (2): 325–327. https://doi.org/10.1007/BF02196732.

Joseph, Jonathan, and Colin Wight, eds. 2010. *Scientific realism and international relations.* Palgrave Macmillan.

Kampen, Jarl K. 2020. "A proposal for the demarcation of theory and knowledge." *Metaphilosophy* 51 (1): 97–110. https://doi.org/10.1111/meta.12398.

Kaur, Arshdeep, and Bojan Božzić. 2020. "Convolutional neural network-based automatic prediction of judgments of the European Court of Human Rights." *27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, CEUR Workshop Proceedings* 2563:458–469.

Kissinger, Henry A. 2018. *How the Enlightenment ends: Philosophically, intellectually — in every way — human society is unprepared for the rise of artificial intelligence.* Technology. The Atlantic, June. Accessed March 1, 2023. https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

Kissinger, Henry A., Eric Schmidt, and Daniel Huttenlocher. 2019. *Metamorphosis: AI will bring many wonders. It may also destabilize everything from nuclear détente to human friendships. We need to think much harder about how to adapt.* Technology. The Atlantic, August. Accessed March 20, 2023. https://www.theatlantic.com/magazine/archive/2019/08/henry-kissinger-the-metamorphosis-ai/592771/.

———. 2021. *The age of AI: And our human future.* Little, Brown and Company.

Kitcher, Philip S. 2023. *Philosophy of science.* Online ed. Encyclopedia Britannica, March 9, 2023. Accessed April 1, 2023. https://www.britannica.com/topic/philosophy-of-science.

Koziol, Helmut, ed. 2015. *Basic questions of tort law from a comparative perspective.* Civil law. Jan Sramek Verlag. https://doi.org/10.26530/oapen_574832.

Lakens, Daniel. 2017. *Impossibly hungry judges.* The 20% Statistician (blog), July 3, 2017. Accessed March 25, 2023. http://daniellakens.blogspot.com/2017/07/impossibly-hungry-judges.html.

Lamond, Grant. 2016. "Precedent and analogy in legal reasoning." In *The Stanford Encyclopedia of Philosophy*, Spring 2016, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Lavrysen, Laurens. 2018. "Causation and positive obligations under the European Convention on Human Rights: A reply to Vladislava Stoyanova." *Human Rights Law Review* 18 (4): 705–718. https://doi.org/10.1093/hrlr/ngy027.

Leiter, Brian, and Michael Sevel. 2022. *Philosophy of law.* Online ed. Revised and updated by Jeannette L. Nolen. Encyclopedia Britannica, August 9, 2022. Accessed April 1, 2023. https://www.britannica.com/topic/philosophy-of-law.

Letsas, George. 2007. *A theory of interpretation of the European Convention on Human Rights.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199203437.001.0001.

———. 2013. "The ECHR as a living instrument: Its meaning and legitimacy." Chap. 4 in *Constituting Europe: The European Court of Human Rights in a national, European and global context*, edited by Andreas Føllesdal, Birgit Peters, and Geir Ulfstein, 106–141. Studies on Human Rights Conventions. Cambridge University Press. https://doi.org/10.1017/CBO9781139169295.005.

Lifschitz, Vladimir. 2019. *Answer set programming.* Springer. https://doi.org/10.1007/978-3-030-24658-7.

Linant de Bellefonds, Xavier. 1994. "L'utilisation d'un "système expert" en droit comparé." *Revue internationale de droit comparé* 46 (2): 703–718. https://doi.org/10.3406/ridc.1994.4899.

Liu, Xinghan, Emiliano Lorini, Antonino Rotolo, and Giovanni Sartor. 2022. "Modelling and explaining legal case-based reasoners through classifiers." In *Legal knowledge and information systems*, E-book, edited by Enrico Francesconi, Georg Borges, and Christoph Sorge, 362:83–92. Frontiers in artificial intelligence and applications. IOS Press. https://doi.org/10.3233/FAIA220451.

MacCormick, Neil. 1992. "Legal deduction, legal predicates and expert systems." *International Journal for the Semiotics of Law* 5 (2): 181–202. https://doi.org/10.1007/BF01101868.

Mak, C. 2012. "Hedgehogs in Luxembourg? A Dworkinian reading of the CJEU's case law on principles of private law and some doubts of the fox." *European Review of Private Law* 20 (2).

M̈aki, Uskali. 2016. "Philosophy of interdisciplinarity: What? Why? How?" *European Journal for Philosophy of Science* 6 (3): 327–342. https://doi.org/10.1007/s13194-016-0162-0.

McNamara, Paul, and Frederik van de Putte. 2022. "Deontic logic." In *The Stanford Encyclopedia of Philosophy*, Fall 2022, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Medvedeva, Masha, Xiao Xu, Martijn Wieling, and Michel Vols. 2020. "JURI SAYS: An automatic judgement prediction system for the European Court of Human Rights." Edited by Serena Villata, Jakub Harašta, and Petr Křemen. *Legal Knowledge and Information Systems: JURIX 2020: The 33rd Annual Conference* (Brno, Czech Republic), 277–280.

Meester, Ronald, and Klaas Slooten. 2021. *Probability and forensic evidence: Theory, philosophy, and applications.* Cambridge University Press. https://doi.org/10.1017/9781108596176.

Moore, Michael S. 2009. *Causation and responsibility: An essay in law, morals, and metaphysics.* Oxford university Press. https://doi.org/10.1093/acprof:oso/9780199256860.001.0001.

———. 2019. "Causation in the law." In *The Stanford Encyclopedia of Philosophy*, Winter 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Moreira, Nídia Andrade. 2022. "The Compatibility of AI in Criminal System with the ECHR and ECtHR Jurisprudence." In *Progress in Artificial Intelligence*, edited by Goreti Marreiros, Bruno Martins, Ana Paiva, Bernardete Ribeiro, and Alberto Sardinha, 108–118. Cham: Springer International Publishing.

Morgan, R. M. 2018. "Forensic science needs both the 'hedgehog' and the 'fox'." *Forensic Science International* 292:e10–e12. https://doi.org/10.1016/j.forsciint.2018.08.026.

Morris, Jason. 2021. "Dynamics of judicial Answer Set Programming as a tool to improve legislative drafting: A rules as code experiment." In *ICAIL '21: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 262–263. Association for Computing Machinery. https://doi.org/10.1145/3462757.3466084.

MSI-AUT (CoE's committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence), Rapporteur: Karen Yeung. 2019. *Responsibility and AI: A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework.* Council of Europe study. DGI(2019)05. Printed at the Council of Europe.

Navarro, Pablo E., and Jorge L. Rodríguez. 2014. *Deontic logic and legal systems.* Cambridge Introductions to Philosophy and Law. Cambridge University Press. https://doi.org/10.1017/CBO9781139032711.

Neelakantan, Arvind. 2017. "Knowledge representation and reasoning with deep neural networks." PhD diss., University of Massachusetts Amherst (College of Information and Computer Sciences), November.

Neves, Marcelo. 2021. *Constitutionalism and the paradox of principles and rules: Between the Hydra and Hercules.* Oxford University Press. https://doi.org/10.1093/oso/9780192898746.001.0001.

Nitta, Katsumi, and Ken Satoh. 2020. "AI applications to the law domain in Japan." *Asian Journal of Law and Society* 7 (3): 471–494. https://doi.org/10.1017/als.2020.35.

Nussberger, Angelika. 2020. *The European Court of Human Rights.* 1st ed. (online). Edited by Mark Janis, Douglas Guilfoyle, Stephan Schill, Bruno Simma, and Kimberley Trapp. Elements of International Law. Oxford University Press. https://doi.org/10.1093/law/9780198849643.001.0001.

O'Shaughnessy, Matthew, Gregory Canal, Marissa Connor, Mark Davenport, and Christopher Rozell. 2020. "Generative causal explanations of black-box classifiers." In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 5453–5467. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc.

OECD's Directorate for education and skills. 2021. *OECD Centre for Educational Research and Innovation (CERI).* Information brochure. Accessed March 25, 2023. https://www.oecd.org/education/ceri/brochure.pdf.

Palmer, Frank R. 2001. *Mood and modality.* 2nd ed. Cambridge textbooks in linguistics. Cambridge University Press.

Parthey, Heinrich. 1999. "Interdisziplinarität – herausforderungen an die wissenschaftlerinnen und wissenschaftler: Festschrift zum 60. Geburtstag von Heinrich Parthey." Edited by Walther Umstätter and Karl F. Wessel, 243–254.

Plakokefalos, Ilias. 2015. "Causation in the law of state responsibility and the problem of overdetermination: In search of clarity." *The European Journal of International Law* 26 (2): 471–492. https://doi.org/10.1093/ejil/chv023.

Poggi, Francesca. 2021. "Defeasibility, law, and argumentation: A critical view from an interpretative standpoint." *Argumentation* 35 (3): 409–434. https://doi.org/10.1007/s10503-020-09544-w.

Prakken, Hendrik (Henry). 1993. "Logical tools for modelling legal argument." PhD diss., Vrije Universiteit.

Priest, Graham. 2014. *One: Being an investigation into the unity of reality and of its parts, including the singular object which is nothingness.* Oxford University Press.

Rawls, John. 1999. *A theory of justice.* Revised ed. Belknap Press.

———. 2000. *The law of peoples with "The idea of public reason revisited".* Revisited. Harvard University Press.

Raz, Joseph. 1979. *The authority of law: Essays on law and morality.* Oxford University Press.

Reiss, Julian, and Jan Sprenger. 2020. "Scientific objectivity." In *The Stanford Encyclopedia of Philosophy*, Winter 2020, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Rigoni, Adam W. 2014. "Legal rules, legal reasoning, and nonmonotonic logic." PhD diss., University of Michigan.

Robins, Sarah, John Symons, and Paco Calvo, eds. 2020. *The Routledge companion to philosophy of psychology.* 2nd ed. Routledge Philosophy Companions. Routledge.

Roth, Bram. 2003. "Case-based reasoning in the law: A formal theory of reasoning by case comparison." PhD diss., Universiteit Maastricht. 10.26481/dis.20031126ar.

Royakkers, Lambèr M. M. 1998. *Extending deontic logic for the formalisation of legal rules.* Edited by Francisco Laporta, Aleksander Peczenik, and Frederick Schauer. Law and Philosophy Library. Kluwer Academic Publishers.

Russell, Stuart J., Peter Norvig, Ming-Wei Chang, Jacob Devlin, Anca Dragan, David Forsyth, Ian Goodfellow, et al. 2021. *Artificial Intelligence: A modern approach.* 4th ed. Gobal ed. Edited by Stuart Russell and Peter Norvig. Pearson series in artificial intelligence. Pearson.

Russo, Federica. 2022. *Techno-scientific practices: An informational approach.* Rowman & Littlefield Publishers.

Sartor, Giovanni. 2009. "Understanding and applying legal concepts: an inquiry on inferential meaning." In *Concepts in law*, edited by Jaap C. Hage and Dietmar von der Pfordten, 88:35–54. Law and Philosophy Library. Springer.

———. 2012. "Defeasibility in legal reasoning." Chap. 6 in *The logic of legal requirements: Essays on defeasibility*, edited by Beltrán Jordi Ferrer and Ratti Giovanni Battista, Part I: General features of defeasibility in law and logic, 108–136. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199661640.003.0007.

Sartor, Giovanni, and Andrea Loreggia. 2020. *The impact of algorithms for online content filtering or moderation - upload filters.* Study requested by the JURI Committee. European Parliament Think Tank. Accessed November 1, 2022. https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2020)657101.

Schaffer, Jonathan. 2000. "Trumping preemption." *The Journal of Philosophy* 97 (4).

Schauer, Frederick. 1991. *Playing by the rules: A philosophical investigation of rule-based decision-making in law and life.* Clarendon Law Series. Oxford University Press.

Schmidt, Jan C. 2022. *Philosophy of interdisciplinarity: Studies in science society and sustainability.* Edited by Alfred Nordmann. History and Philosophy of Technoscience. Routledge.

Schroeter, Francois, Laura Schroeter, and Kevin Toh. 2020. "A new interpretivist metasemantics for fundamental legal disagreements." *Legal Theory* 26 (1): 62–99. https://doi.org/10.1017/S1352325220000063.

Sergot, M. J., F. Sadri, R. A. Kowalski, F. Kriwaczek, P. Hammond, and H. T. Cory. 1986. "The British Nationality Act as a logic program." *Commuications of the Association for Computing Machinery* (New York, NY, USA) 29, no. 5 (May): 370–386. https://doi.org/10.1145/5689.5920.

Shafer, Glenn. 2002. "Causality and responsibility." In *The dynamics of judicial proof: Computation, logic, and common sense*, edited by Marilyn MacCrimmon and Peter Tillers, Part VIII: Causality, 457–478. Studies in fuzziness and soft computing. Physica-Verlag.

Singh, Ajay K. 2010. "ESAs in dialysis patients." *Journal of the American Society of Nephrology* 21 (4): 543–546. https://doi.org/10.1681/ASN.2010020178.

Sivaram, Venkat, Abhishek; Venkatasubramanian. 2022. "XAI-MEG : Combining symbolic AI and machine learning to generate first-principles models and causal explanations." *AIChE Journal* 68 (6). ISSN: 0001-1541. https://doi.org/10.1002/aic.17687. https://browzine.com/articles/520116101.

Solan, Lawrence M., and Peter M. Tiersma, eds. 2012. "The Oxford Handbook of language and law." Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199572120.001.000.

Stavropoulos, Nicos. 1996. *Objectivity in law.* Oxford University Press.

Steffen, Will, Angelina Sanderson, Peter Tyson, Jill Jäger, Pamela Matson, Berrien Moore, Frank Oldfield, et al. 2005. *Global change and the earth system: A Planet Under Pressure.* Global Change - The IGBP Series. Springer. https://doi.org/10.1007/b137870.

Stoyanova, Vladislava. 2018. "Causation between state omission and harm within the framework of positive obligations under the European Convention on Human Rights." *Human Rights Law Review* 18 (2): 309–346. https://doi.org/10.1093/hrlr/ngy004.

Sulyok, Katalin. 2017. "Managing uncertain causation in toxic exposure cases: Lessons for the European Court of Human Rights from U.S. toxic tort litigation." *Vermont Journal of Environmental Law* 18:519–569.

Taylor, Luke. 2023. "Colombian judge says he used ChatGPT in ruling: Juan Manuel Padilla asked the AI tool how laws applied in case of autistic boy's medical funding, while also using precedent to support his decision." (Bogotá) (February 3, 2023). Accessed March 11, 2023. https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling.

Tegmark, Max. 2018. *Life 3.0: Being human in the age of artificial intelligence.* Penguin.

Thorén, Henrik, and Johannes Persson. 2013. "The philosophy of interdisciplinarity: Sustainability science and problem-feeding." *Journal for General Philosophy of Science* 44 (2): 337–355. https://doi.org/10.1007/s10838-013-9233-5.

Timmer, Ivar, and Rachel Rietveld. 2019. "Rule-based systems for decision support and decision-making in Dutch legal practice." *Droit et société* 103 (3): 517–534.

Towell, Geoffrey G., and Jude W. Shavlik. 1994. "Knowledge-based artificial neural networks." *Artificial Intelligence* 70 (1): 119–165. https://doi.org/10.1016/0004-3702(94)90105-8.

Turton, Gemma. 2020. "Causation and risk in negligence and human rights law." *The Cambridge Law Journal* 79 (1): 148–176. https://doi.org/10.1017/S0008197319000898.

Ulenaers, Jasper. 2020. "The impact of artificial intelligence on the right to a fair trial: Towards a robot judge?" *Asian Journal of Law and Economics* 11 (2): 20200008. https://doi.org/doi:10.1515/ajle-2020-0008.

van Woerkom, Wijnand, Davide Grossi, Henry Prakken, and Bart Verheij. 2022. "Landmarks in case-based reasoning: From theory to data." In *HHAI2022: Augmenting human intellect*, E-book, edited by Stefan Schlobach, María Pérez-Ortiz, and Myrthe Tielman, vol. 354. Frontiers in artificial intelligence and applications. IOS Press.

Vann, Richard T. 2023. *Intellectual history.* Online ed. Encyclopedia Britannica, March 17, 2023. Accessed March 24, 2023. https://www.britannica.com/topic/intellectual-history.

Vaquero, Álvaro Núñez. 2013. "Five models of legal science." Translated by Ester González Bertrán. *Revus* 19. https://doi.org/https://doi.org/10.4000/revus.2449.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention Is All You Need.* arXiv: 1706.03762 [cs.CL].

von der Lieth Gardner, Anne. 1987. *An artificial intelligence approach to legal reasoning.* Edited by L. Thorne McCarty and Edwina L. Rissland. Artificial intelligence and legal reasoning. The MIT Press.

von Wright, Georg Henrik. 1951. "Deontic logic." *Mind* 60:1–15.

Waldron, Jeremy. 2020. "The rule of law." In *The Stanford Encyclopedia of Philosophy,* Summer 2020, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Walton, Douglas. 2002. *Legal argumentation and evidence.* The Pennsylvania State University Press.

Waluchow, Wil, and Stefan Sciaraffa, eds. 2016. *The legacy of Ronald Dworkin.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190466411.001.0001.

Wan, Hui, Michael Kifer, and Benjamin Grosof. 2015. "Defeasibility in answer set programs with defaults and argumentation rules." *Semantic Web* 6:81–98. https://doi.org/10.3233/SW-140140.

Winter, Jack. 2016. "Justice for hedgehogs, conceptual authenticity for foxes: Ronald Dworkin on value conflicts." *Res Publica* 22 (4): 463–479. https://doi.org/10.1007/s11158-015-9285-y.

Wright, Richard W. 1985. "Causation in tort law." *California Law Review* 73:1735–1828.

———. 1988. "Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts." *Iowa Law Review* 73:1001–1077.

———. 2011. "The NESS account of natural causation: A response to criticisms." Chap. 14 in *Perspectives on causation,* edited by Richard Goldberg, 285–322. Hart Publishing.

Yang, Zhun, Adam Ishay, and Joohyung Lee. 2020. "NeurASP: Embracing neural networks into Answer Set Programming." In *Proceedings of the twenty-ninth international joint conference on artificial intelligence,* edited by Christian Bessiere, 1755–1762. International joint conferences on artificial intelligence organization. https://doi.org/10.24963/ijcai.2020/243.

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. "Hierarchical attention networks for document classification." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1480–1489. San Diego, California: Association for Computational Linguistics, June. https://doi.org/10.18653/v1/N16-1174.

Yoshino, Hajime. 1987. "Legal expert system LES-2." In *Springer lecture notes in computer science presented at logic programming' 86, Tokyo, 23–26 June 1986,* 34–45. Springer.

———. 1998. "Logical structure of contract law system—for constructing a knowledge base of the United Nations Convention on Contracts for the International Sale of Goods." *Journal of Advanced Computational Intelligence and Intelligent Informatics* 2:2–11.

CHAPTER III

Gluing
The art of going META

In CHAPTER I, I argued that for a legal ALGOAI model to be *legitimate*, it needs to provide *justifications* for its output *similar* to the normative rational justifications human authorities are expected to provide. In CHAPTERS I & II, I further argued that it is the job of the logician & the formal philosopher to be the fox that glues the different ALGOAI engineering disciplinary practices in order to *identify & formalise* the normative reasoning methods of human judicial authority so as to incorporate them into legitimate legal ALGOAI models. The *objective* of this chapter is to provide the *foundations* for a *methodology* for performing such gluing. I.e., the foundations for engineering *legitimate formal models of judicial justifications*.

More precisely, in §1, I introduce the concept of *(semi-)formal model* since I intent to engineer a semi-formal model of judicial justifications. In §2, I delineate the qualities that a *method* for engineering (semi-)formal models should have. I further argue why Carnap's conceptual re-engineering method of *explication* is a method that has those qualities. In §3, I adjust Carnap's explication so as to model (semi-)formal models of judicial justifications that are intended to be incorporated into legitimate legal ALGOAI. Finally, in §3.3, I argue how the practice of explicating should start: which are the first steps that the explication engineers should follow? Throughout all those sections, I am using as a toy example of judicial justifications *causal* justifications employed by the ECtHR.

In the next chapter, CHAPTER IV, I apply the methodology of explication I lay out in this chapter to model ECtHR's causal justification. The goal is not to provide a full-fledged model. After all, that would be impossible since I lack the expertise and resources to do so. It would need years, millions (if not billions) of euros, and a CROSSDI team of top experts from AI, legal science, logic, formal philosophy, and other disciplines so as to engineer a model adequate enough to be used in real life. My goal is rather to exhibit *how* the foundations of explication introduced in this chapter can be used in the ALGOAI engineering practice.

III.1 What is a model?

I stated that the objective of this chapter is to draft a methodology for designing a specific *object*, that object being a *model of judicial justifications*. What constitutes though an *understanding* of an object? Borrowing from philosophy of explanation and cognitive science, two ways of understanding an object are: (a) identifying the *constitutive elements* that are glued together to form the object (e.g., Siscoe 2022); (b) identifying the *uses* of the object (e.g., Frost and Monaghan 2020; Speaks 2021, §§3.2.4-3.2.5). I chose those two ways of understanding for practical reasons. Specifying the constitutive elements of the model will allow me to break my methodology into *steps* with each step addressing different constitutive elements. Specifying the use of the model will offer decisive guidance for determining the *content* of those steps; I will choose content that will be of practical importance.

a model's *constitutive elements*

Luckily, Smaldino 2017 (pp.313-315) provides a conceptualisation of models that answers both requirements of understanding. The *constitutive elements* of a model are *parts* (or *relata*), *relations* among those parts, and *properties* that the relata and the relations have. Look for instance Figure 1:

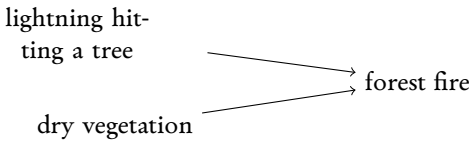


Figure 1: This is a toy model of causal inference. It is a variation of Figure 2.1 (b) in Halpern 2016, p.16.

The model's *parts* are the phrases "*lightning hitting a tree*", "*dry vegetation*", and "*forest fire*". Those words represent actual or hypothetical *instances* of lightnings hitting trees, dry vegetation, and forest fires. The *relations* among the parts of the model are the arrows "$\longrightarrow$" that represent causal relations: in the tail of the arrow, we have the cause, and in its head, we have the cause's effect. The arrow's one-directionality represents the property of *asymmetry* that the causal relation has: the causes cause the effect but not the other way round (*ibid.*, p.17). Note that the constitutive elements of the model of Figure 1 are the symbols and not what the symbols represent (e.g., the phrase "*dry vegetation*" and not actual or hypothetical instances of dry vegetation) (*cf.* §I.3.2.1.1, ¶9). This construal of the Figure 1's model is *on par* with the view that models are *representations* of a *state of the world*, "*the received view in philosophy of science*" (Russo 2022, §5.1.2).

But what is the *use* of models? Smaldino 2017 argues that the use of a model is to answer a specific *research question*. The research question of Figure 1 can be "*How are forest fires caused?*". The research question of our model is "*How does the ECtHR causally justify its judgments?*". In order to answer the research question, we will have to answer further more precise *subquestions*. However, those are not and can not always be decided *ex ante*. Many of them will be fleshed out while we specify constitutive elements of the model. Identifying the parts of the model, their relations, and their respective properties allows us to have a language to articulate new questions and precisify old ones (pp.314-315). The formulation of the research question is either motivated by a genuine academic interest in the question itself (e.g., we are academically interested in how the ECtHR causally justifies its judgments). Or it is a proxy for achieving TRANSDI ends (e.g., to be used in ALGOAI judicial models). I will name the answer to the research question as the *purpose* of the model and the TRANSDI ends in case there are such as the *intended application* of the model. Smaldino 2017 fails to make such a distinction ending up conflating the two dimensions of use and not adequately grasping the TRANSDI use of models.Why is that distinction important though?

a model's *use:* purpose *v.* intended application

You see, models are, as Smaldino 2017 phrases it, *stupid* (sic). And they *should* be stupid. Identifying the TRANSDI ends of the intended application is important because those ends are what guide us to stupid down the model to the optimal degree of stupidity. In this context, "*stupidity*" means that the model does not capture every aspect of reality that contributes to answering the research question, i.e., to fulfilling the purpose of the model. The model is and it should be a *simplification* of reality. This is because reality is very complex and answering the research question to its full extent would be impractical, if not unattainable. Lucky for us, we care about the aspects of the research question which are of relevance to the intended application, and hence, we can disregard the rest of reality (*cf.* §IV.1, ¶2). For instance, in the real world, the reason why a forest fire starts is more complicated than what is depicted in Figure 1. Apart from lightning hitting a tree and the dry vegetation of the forest, it may also be the case that the local authorities did not take sufficient fire prevention measures. Or that the firefighters were late or understaffed. All of those aspects of reality are part of the answer to the research question of how the fire started. However, if we intend to use the model for identifying the contribution of *physical* phenomena to the forest fire, we do not have to include in the model the negligence of the local authorities or the response of the firefighters.

why models *should* be *stupid*

In the case of ALGOAI engineering, apart from disregarding the *noise* of the irrelevant information, there is another reason for wanting to reduce the *complexity* of the model. Part of the art of AI modelling is deciding which parts of reality should be left out of the model in order to make the AI operate in a reasonable time (more on §3.2.3) which is *inter alia* a rule of law requirement (§I.2.4, ¶8). Ergo, legal ALGOAI engineers are required to find the *optimal* balance between maximizing the model's usefulness and minimizing its complexity. And it is the intended application and not the purpose of the model that will guide them to separate efficiently the wheat from the chaff.

stupidity *v.* complexity

Now that I explained what is a model, it is time to explain what is a *formal* model

III.1.1 What is a *formal* model?

We saw in §II.4.1.1, ¶4 that in order to *design*, *build*, and *evaluate* a model, we need *languages* $\mathcal{L}_d$, $\mathcal{L}_b$, and $\mathcal{L}_e$ respectively. We can categorise models based on the language $\mathcal{L}_b$. For instance, *verbal* models are those for

formal models

which we use natural languages, while *formal* models are those for which we use formal languages (Smaldino 2017, pp.4-6). As formal languages, I construe the languages of mathematics, logic and computer science.[1]

We do not have to make a strict distinction between formal models and other kinds of models; we can also have *hybrid* models. I will call such models *semi-formal*. Our model will be semi-formal. Since in the CROSSDI ALGOAI engineering team there will most likely be experts that are not that familiar with formal languages, it is essential to include non-formal tools that will make the model *epistemically accessible* to those experts. More importantly, since the engineered models are intended to be used by judicial authorities and their subjects, the use of epistemically accessible language is a *necessary* legitimacy requirement (§I.2.4, ¶¶5-9).

semi-formal models

III.2 In the search of a methodology

III.2.1 What makes a methodology *good*?

Summing up §1, the objective of this chapter is to conceptualise a *methodology* for engineering a semi-formal model. It is a case of *methodology*-oriented TRANSDI CROSSDI.[2] How will I know though that my methodology will be a *good* methodology? What does even mean for such a methodology to be "*good*"? Clearly, I need some criteria of goodness. I construe a *good* methodology for engineering a model to be a methodology that can produce *good* models. I construe a *good* model to be a model that *sufficiently* fulfills its purpose and that *sufficiently* realises its intended application.[3] The *evaluation criteria* of that sufficiency should not be chosen by me; I can always choose criteria that make my model be good if I am the one determining goodness. I need to employ a well-established in the literature framework of evaluation that is accepted by the experts as a benchmark. Which are those *experts* though? Is there a community of experts whose expertise is about methodologies for designing semi-formal models?

The answer is *yes*, there is! One such community is *formal philosophers* showcasing once more the importance of their contribution to an ALGOAI engineering team. Formal philosophy is the practice of philosophy with the employment of *formal methods*. I.e., methods from mathematics, logic, and computer science (e.g., formal ethics, formal epistemology) (Leitgeb 2013, §2).[4] Since formal methods make use of formal languages, formal philosophy can fruitfully interact with disciplines that also employ formal languages. For instance, discussing which *criteria* make a scientific output a *good* output is a classical task of formal philosophy of science. Those criteria can be used by actual scientists in their practice allowing a fruitful interaction among the disciplines. We will see in CHAPTER IV examples of such fruitful interactions with the use of formal models of causal justifications by disciplines like law, computer *&* cognitive science (*fn.* 23). Considering the above, to engineer a model of causal justifications, I will use a methodology of engineering models developed by formal philosophers. Which one of the available methodologies should I choose though?

A category of such methodologies is the *conceptual re-engineering* methodologies. Conceptual re-engineering is the process of substituting an already existing *concept* with a *better* version of that concept (Chalmers 2020, ¶1, p.6). By *substitution*, one means that in a certain context, the interlocutors will use the re-engineered concept instead of the original one. Note that the characterisation "*conceptual re-engineering*" can be attributed retrospectively to a concept substitution without those performing the substitution being aware that what they are doing is conceptual re-engineering. For instance, two decades before the term "*conceptual engineering*" was coined, Carnap introduces what is now considered a paradigmatic example of conceptual re-engineering: *explication* (Isaac 2020, p.4, *fn.* 4; Chalmers 2020, p.4). A typical example of explication is the substitution the concept of *warmness* by the *better* concept of *temperature* (Carnap 1962, §I). E.g., for meteorologists, saying merely that "*the weather will be warm*" is not enough. They have to provide an exact temperature: "*the weather will be* $28°C$".[5,6] What makes a concept *better* than another one depends on the *ends* of each re-engineering

conceptual re-engineering

[1] By "*language*", I do not refer to any formal definition of languages like those from formal linguistics or logic. Having said that, such formal definitions can be used to design formal models since they belong to formal languages. Moreover, in the language of computer science, I include algorithmic processes (Smaldino 2017, ¶3,p.5) as well as programming languages.

[2] Remember though that the *methodology*-oriented direction will inevitably alternate among *object*-oriented *&* *theory*-oriented directions at some point (§II.3.1.2.1, ¶¶6-7).

[3] I say "*sufficiently*" since on the one hand, there must be a minimum standard of *goodness* that the model meets, while on the other hand, it would be arrogant to claim that one can design a model for which there is no room for improvement.

[4] Leitgeb differentiates formal methods from methods of computer science (what he calls "*computational methods*") (Leitgeb 2013, §2). Personally, I construe formal methods as methods that employ *formal languages*, and hence, formal philosophy includes computational methods. However, I do acknowledge that a methodology is more than the language one uses and hence Leitgeb's distinction is more accurate. For practical reasons though, I will use only *one* umbrella term: "*formal methods*".

[5] Carnap's arguments of why the substitution of the concept of warmness from the concept of temperature is an instance of explication and why the latter is a "*better*" concept can be found in Carnap 1962, §§I.4-I.5.

[6] Examples of conceptual re-engineering in *law* are Rawls' 1999 re-engineering of the concepts of *justice* and *rightness* in his seminal "*A theory of justice*", Griffin's 2008 re-engineering of the concept of *human rights* (Brun 2020), Spaak's 2009 "*Explicating the concept of legal*

method. For example, Carnap's explication aims at *semantical clarity*. Thus, Carnap argues that a criterion of betterness should be whether the re-engineered concept is more *exact* like the temperature example (Carnap 1963, pp.935-936). On the other hand, for the re-engineering method of *ameliorative analysis* introduced by Haslanger 2012, semantical clarity is not enough. The re-engineered concept needs to serve certain *political* ends (Dutilh Novaes 2020, p.1026). E.g., re-engineering the concepts of *race* or *gender* so as to "*[fight] inequality*" (*ibid.*, p.1023).

In the following subsection (§2.2), I argue that *engineering a mode*l of the ECtHR's causal justifications is in reality a *conceptual re-engineering* of the concept of causal justification as it used in the ECtHR's practice. Based on that, I argue that I can use Carnapian *explication*'s criteria of *betterness* to evaluate whether a model of causal justification is a *good* model. And consequently, whether the methodology I use to engineer that model is a *good methodology*. In the literature, the explication's criteria of *betterness* are referred to as criteria of *adequacy* (*see e.g.* Brun 2016, §2.3; 2020, §3.3; Dutilh Novaes 2020, §3.1; Isaac 2020, pp.4,9). Therefore, instead of saying "*better*", "*good*", "*sufficiently*", I will be saying "*adequate*" and "*adequately*": adequate model, adequate methodology, adequate concept, adequately fulfilling the model's purpose, adequately realising the model's intended application.

III.2.2 Modelling as conceptual re-engineering

To argue that modelling ECtHR's causal justification is a re-engineering of the concept of the ECtHR's causal justification, I will have to argue that: (a) ECtHR's causal justification is a *concept*; (b) modelling that concept is a *re-engineering* of that concept. This is what I do in the next two subsections respectively.

III.2.2.1 Concepts, concept-hood, & causal justification

Let's begin by arguing that ECtHR's methodology of causal justification is a *concept*. There are multiple conflicting theories as to what concepts are. The examples of explication and ameliorative analysis were about re-engineering a concept so as to achieve among others *semantical* clarity. That places them in the philosophical tradition that construes concepts as the "*meaning of words and phrases*" (Margolis and Laurence 2022, §1.3). Indeed, if one looks at the literature on conceptual re-engineering (*see e.g.* the paradigmatic cases of conceptual engineering in Chalmers 2020, pp.4-5), the re-engineering has to do with *semantics* of words or phrases: how can the meaning of the terms justice, rightness, human rights, mineralogical hardness, fish, gender, woman, consciousness, rigid designator, truth, supervenience, probability, cold, numbers, opinion, belief, explanation, and certainty be *improved*?[7]

the *semantic* construal of concepts

Before moving forward, let's standardise my terminology based on the conception of concepts as *meanings*. To do so, we would have to adopt a specific theory of meaning. Carnap does not do so trying to remain as neutral as possible. Let's try to be equally practical. Brun notices that Carnap does not adopt a specific theory of meaning and instead they focus on the *practical* aspect of re-engineering as a method of improving a way that a concept is used. More precisely, they propose an interpretation of Carnap's explication according to which a concept is: (a) the *term* (word or phrase) used to denote that concept (e.g., "*temperature*" is the term used to denote the concept of temperature); (b) the term's *rules of use*.[8] Consequently, two concepts are *identical if and only if* both their terms and rules of use are identical (Brun 2016, pp.1216-1217). Re-engineering is the process in which we *improve* the rules of use where improvement is construed in terms of specific *adequacy criteria*. Sometimes the *term* of the concept also changes (e.g., "*warm*" changing to "*temperature*"). Sometimes it does not (e.g., Rawls' re-engineered concept of justice makes use of the same term "*justice*"). To denote the term of a concept I will use double quotation marks and *italics* (e.g., "*causal justification*"). To denote a concept I

what is a concept?

competence". For the hardcore Rawlsians, they can also compare Rawls' re-engineering of the 1999 revised edition of "*A theory of justice*" with the re-engineering of the original 1971 edition. The logical model of the interpretation & application of the law by judicial authorities introduced in §II.4.1.2 is also a case of conceptual re-engineering.

[7] An underrepresented alternative in conceptual engineering is the construal of concepts as *cognitive abilities* (Margolis and Laurence 2022, §1.2). For instance, the concept of dog is not the meaning of the word "*dog*", but it is our cognitive ability to make inferences about dogs based on a body of information regarding dogs (Machery 2017, p.210). Isaac 2020 attempts to conceptually re-engineer the concept of *conceptual engineering* so as to be about the cognitive construal of concepts and not the semantical one. Hence, they call it *cognitive engineering*. For a similar cognitive approach *see* Prinzing 2018. Carnap 1962 (p.8) explicitly rejects any subjective construal of concepts as they are being cognitively conceived instead of being the objective meanings of words/phrases.

[8] Brun stays neutral as to whether a term's rules of use are about the *extension* or the *intention* of the term. As we will see in §3.2.1, I introduce rules of use for both. Brun also stays neutral as to whether a term's rules of use can be related to other conceptions of concepts like *mental representations* (Brun 2016, p.1217). As Margolis and Laurence note, there are philosophical theories that combine different conceptions of concepts (e.g., semantical and cognitive conceptions of concepts) (Margolis and Laurence 2022, §1.3). Hence, there might be rules of use that influence all those conceptions of concepts. The fact that Brun's interpretation of Carnap is compatible with such different philosophical positions, along with the fact that his 2016 paper is considered a seminal work on Carnap's explication (*see e.g.* Chalmers 2020, p.6) are *the two reasons for which I follow his approach*.

will use SMALL CAPS (e.g., CAUSAL JUSIFICATION).[9] The concept that is explicated is called *explicandum* (plural *explicanda*) and the re-engineered concept that explicates it is called *explicatum* (plural *explicata*) (Carnap 1962, ¶2, p.3).

Considering the above, CAUSAL JUSTIFICATION *is* a concept. Its *term* is "*causal explanation*" and its *rules of use* are those that are implicitly and explicitly found in the ECtHR's legal tradition of human rights law. In what follows, I argue that *modelling* a normative model of CAUSAL JUSTIFICATION is in reality modelling a more adequate version of its rules of use, and ergo, it is a case of conceptual re-engineering and in particular of the Carnapian explication. Before doing so though, I would like to address two other aspects of the *objectivity challenge* that conceptual engineers will inevitably face.

III.2.2.1.1 Objectivity challenge again

Since a concept is a *term* and its *rules of use*, the same term but different rules of use denote different concepts. For instance, the concept CAUSATION has different rules of use in different legal traditions & areas of law in each legal tradition (Moore 2019, §1; §I.1.1, ¶2). Consequently, when we use the term "*causation*", we end up with another ambiguity of reference: to which of all these concepts of causation we are referring? The objectivity challenge strikes again! It's time to bite the rest of the bullet.

We saw at the end of CHAPTER I, that the principles of foreseeability and legality resolve a significant part of the objectivity challenge. The *ordo essendi* that the ALGOAI engineers will use is the one that emerges in the practice of the judicial authority that the ALGOAI model will support or replace (§I.3.2.1.2, ¶¶11-12). Choosing a specific *ordo essendi* is essentially designating what Brun 2016 calls *system of concepts*. Choosing systems of concepts is of prime importance for the conceptual engineering method of *explication* since the original concept and the re-engineered concept need to belong to *two distinct* systems of concepts (§4.1, pp.1229-1230). I will call the *explicandum*'s system of concepts *source system* and the *explicatum*'s system of concepts *target system*.[10] Brun does not provide an exact definition of a system of concepts neither in the introduction of the term in Brun 2016 nor in his later work on explication (Brun 2020). Following Enlightenment's paradigm (§I.2.3), I will construe *mechanistically* a system of concepts as a specific collection of concepts and the way in which those concepts are related to each other. In different discourses (e.g., in everyday conversations, in the practice of the ECtHR, in thermodynamics), the interlocutors make use of different systems of concepts. By designating a specific system of concepts, we clarify that we are using concepts *only* from that system.

source & target systems of concepts

Even if we designate though which are the source & target systems of concepts, the objectivity challenge still persists! The first case of reference ambiguity is when in the same system of concepts, the same term is used according to multiple distinct sets of rules that may even be contradictory. In case that it is done deliberately, it does not constitute a problem since there must be already available tools to differentiate between those two concepts. E.g., by using qualifying adjectives. We will see in §IV.1 that this is the case with CAUSATION: legal experts differentiate between two concepts of causation (FACTUAL CAUSATION *v.* LEGAL CAUSATION) by using the qualifying adjectives "*factual*" and "*legal*" respectively.[11] In case the use of different sets of rules for the same term is *undeliberate*, we saw that the law should facilitate procedures where such ambiguities of reference would be made known to the respective judicial authorities so as to find a solution ((§I.3.2.1.2, ¶14).

more objectivity challenge

The second type of ambiguity of reference is not about those who ambiguously use a term in their practice in the source system, but of us that interpret their practice as *external* observers. As Stanford notes, whenever we make judgements about how a concept is used, we also make *interpretive decisions* about "*past speakers and linguistic communities*" regardless of whether we are aware of those decisions (Stanford 2015, pp.406-407). For instance, when we analyse the ECtHR's judgements, we make decisions about what the authors that crafted those judgements meant, as well as what the concepts they were using meant in their "*linguistic community*". I.e., the source system that we analyse is *not* the source system itself, but our *interpretation* of it. Hence, we should not allow ourselves to be mislead into thinking that our judgements about the source system are "*brute facts*" (*ibid.* p.407).

even more objectivity challenge

This is a common point of criticism against *documentary research*.[12] From which documents are *practically available* to us (is the information in those documents enough for our purposes) to our personal ethical, political

[9] The use of SMALL CAPS to denote concepts is borrowed from Schroeter, Schroeter, and Toh 2020, p.66, *fn.* 3. I adopt this notation since in §3.3, I incorporate Schroeter's, Schroeter's, and Toh's methodology of legal interpretation in my proposal of how explication should be.

[10] Brun 2016, 2020 considers systems of concepts as part of *theories*. Hence, Brun uses "*target theory*" instead of "*target system of concepts*". I opt to differ because by using "*theory*", one may be mislead to assume that the *explicatum* can not be part of ordinary language something that Carnap 1963 (p.935) explicitly rejects. Brun 2016 also explicitly acknowledges the possibility of explicating ordinary language concepts (pp.1216,1237).

[11] Another such case that we have already seen (§I.2.7, ¶6) is the distinction between two types of the concept MARGIN OF APPRECIATION in the ECtHR's legal tradition, distinguished by the qualifying adjectives "*substantive*" and "*structural*".

[12] *Documentary research* is an umbrella term used for research whose output is based on the analysis of *documents*, where documents do not have to be solely text-based official reports, but also, documented personal opinions, visual data like photographs, statistical reports,

and social *biases* that shape the prism with which we filter the information of the available documents, all those are all parameters that mediate between our interpretation of the source system and the actual source system as it is shaped in the actual legal practice (May 2011, pp.215-216). That is something that can never be avoided though. The conceptual engineer by default conceives the explicandum at a *meta*-level, and hence, what they use is their conception of what is conceived. What we can do though is employ experts whose expertise allows them to mitigate any interpretive differences, the legal experts. We will also see in §3 explication methods that can mitigate further those differences.

semantical v. syntactical interpretation

Note that this distinction between the concept itself and its interpretation at a *meta*-level is a problem of *conceptual* interpretation (Prakken 1993, p.14), i.e., a type of *semantical* interpretation. In §3.2.2, we will see a second problem of the same kind, that of *syntactical* interpretation. As *syntactical interpretation*, I construe: (α) the interpretation of the (semi-)formal *syntax* of: (α.i) the facts of a case; (α.ii) of the laws applied to those facts; (α.iii) the inference of those facts & laws; (β) the interpretation of the *reaosning method* used to infer a judgement from the facts of a case & the laws that are applied to those facts (e.g., the subsumptive-deductive reasoning method that we saw in §II.4.1.2) (*cf.* Prakken 1993, p.14). Legal experts are more equipped for resolving problems of semantical interpretation, while logicians & formal philosophers are more equipped for resolving problems of syntactical interpretation.

III.2.2.2 Modelling as conceptual re-engineering

Now that I have established that CAUSAL JUSTIFICATION is indeed a concept, I will argue that the *modelling* of causal justification for the purposes of legal ALGOAI engineering is indeed a *re-engineering* of CAUSAL JUSTIFICATION. My argument is that the model I want to construct can be construed as a set of *improved* rules of use of the concept CAUSAL JUSTIFICATION which will be used in its place in a specific system of concepts. I will further argue why explication is a suitable method of conceptual re-engineering to realise the desired TRANSDI ends without excluding other alternatives that can achieve more adequate re-enginered concepts.

First things first, how can a *model* of causal justifications be construed as a set of *rules of use* of the concept CAUSAL JUSTIFICATION? We have seen that at least some of the constitutive elements of a model should have a *semantical* interpretation. I.e., they should be representations of actual relata, relations and properties of what is modelled (§I.3.2.1.1, ¶14; *cf.* §1, ¶2). In our case, what is modelled is ECtHR's causal justifications. Consequently, constitutive elements of the model should represent constitutive elements of *actual* instances of causal justifications. Look for instance Figure 1, a model of CAUSAL INFERENCE. Its parts are the words "*lightning hitting tree*", "*dry vegetation*", "*forest fire*" that represent *actual* instances of lightning hitting a tree, dry vegetation, and forest fires. Those actual occurrences are the constitutive elements of *actual* causal inferences. I.e., every time one infers that a forest fire was caused by a lightning hitting a tree and dry vegetation, that causal inference is represented by that model. That means that the model of Figure 1 does not model the constitutive elements of only one *single* instance of causal inference, but of a *collection* of instances of causal inference. Those instances of causal inference are *particulars* that are *subsumed* the *concept* CAUSAL INFERENCE: they have the property of *being causal inference*. In other words, the model represents part of the *extension* of CAUSAL INFERENCE. In a similar vein with the example of CAUSAL INFERENCE in Figure 1, the model of CAUSAL JUSTIFICATION we want to engineer should not consist of the constitutive elements of a single instance of causal justification but of *multiple* instances of causal justifications so as to be used in multiple judgements (*cf.* with the requirement of *maximising universality* in §3.2.4). Therefore, the model will be a representation of parts of the *extension* of CAUSAL JUSTIFICATION. Considering the above, the model is a representation of at least one rule of use of CAUSAL JUSTIFICATION: IF an object can be represented by the model, THEN it is part of the extension of CAUSAL JUSTIFICATION. I will call this rule as *rule of extension*.[13]

models as rules of use: the rule of extension

Based on the rule of extension we can derive more rules of use! For instance, in Figure 1, when we know that an object is part of the model's extension due to the rule of extension, we are able to derive new rules for *predicting* future events, *justifying* past events, attributing *responsibility*, etc. All of them are regular *uses* of CAUSAL INFERENCE. For instance, if we know that the vegetation is dry and that there is high probability for a lightning storm, we have a state of affairs which is represented by the model of Figure 1. From the rule of extension, we know that we have an instantiation of causal inference. Hence, we can *predict* with high probability that a forest fire will be caused. Similarly, if we know that a lightning hit a tree but there was no forest fire, according to the rule of extension, we can infer that the vegetation was not dry. If it was dry, then we would have had an instantiation of the model, and hence, we would have had a forest fire which was not the

models as rules of use: more rules of use

etc. For more, *see* May 2011, §8 and Webley 2010, §IV.C. Since the ALGOAI engineers have to indeed process documents (the ECtHR's judgements, the Registry's Guides on the articles of the Convention, judges' dissenting opinions, critical academic literature or reports by unelected expert bodies on the ECtHR's judgements, etc), they do perform documentary research.

[13] Henceforth, I will generally express *rules of use* in an IF-THEN form to be on par with the logical model of interpreting & applying the law introduced in §II.4.1.2.

case. Thus, the model *justifies* why there was not a forest fire.[14] Or if we know that both causes happened but the effect did not happen, then we can infer that an external factor, external to our model, "*broke*" the causal relations of the model. If the causal relations were not "*broken*", then from the rule of extension we know that there would have been a forest fire. Now, if we know that what intervened to "*break*" the causal relations were agents that had the *responsibility* to prevent a forest fire (e.g., firefighters), then we can infer that those agents fulfilled their responsibility. From the above, we can see that from the rule of extension we can infer at least the following rules: (a) IF the causes of an effect happen, THEN the effect will follow (*time-asymmetry rule*); (b) IF a cause happened and the event did not follow, THEN an external factor intervened and "*broke*" the causal relation; (c) IF the effect follows from the conjunction of two causes, one of the causes happened, but the effect did not follow, THEN the second cause did not happen; (d) IF the cause happened but the event did not happen due to the intervention of agents that had the responsibility to stop the effect, THEN those agents fulfilled their responsibility. To generalise, by knowing that some of the constitutive elements of the model have been instantiated, then making hypotheses of what would have happened in case that the whole model had been instantiated, and finally *comparing* the results of these hypotheses to (background) knowledge, we can generate rules of use about the model's constitutive elements. I.e., about what those elements *represent* in the actual world. When those *comparisons* use knowledge from different disciplines (e.g., using knowledge from legal science so as to identify rules of use of CAUSAL INFERENCE) what we have is a paradigmatic example of using *comparative* METADI information to produce *contactual* METADI information (§II.3.1.2).

The rule of use are not exhausted here! The *purpose* of the model is to adequately answer the research question: "*How does the ECtHR causally justify its judgments?*". This is essentially a question of *how* the concept CAUSAL JUSTIFICATION is *used* by the ECtHR in its judgments. Consequently, answering that question requires again to identify *rules of use* of "*causal justification*" and incorporate them in the model. In a similar vein, fulfilling adequately the *intended application* of the model means that the model can be employed in AI to generate causal justifications. For the AI to *use* the model to generate causal justifications, we should first know how the ECtHR *uses* "*causal justification*" and then build those rules of use in the model.

models as rules of use: even more rules of use!

So far, I have argued that modelling CAUSAL JUSTIFICATION can be construed as making explicit in a semi-formal language rules of how a concept is used. In conceptual re-engineering though, we make explicit *improved* rules of use. Consequently, to argue that my modelling approach of CAUSAL JUSTIFICATION is a re-engineering of CASUAL JUSTIFICATION, I have to argue that the rules of use of the model are an *improved* version of the rules of use of the original concept. Let's see why this is the case. In §I.3.2.1.2, we saw that by requiring the model to *describe* how the Court uses CAUSAL JUSTIFICATION, we are able to track logical inconsistencies in the Court's use of that concept. Hence, by being descriptive, the model allows us to identify misuses of that concept and consequently to have a more *exact* understanding of its rules of use (*exactness*).[15] Also, by describing the rules of use of CAUSAL JUSTIFICATION using formal language, we are able to incorporate them in other disciplines (AI, formal philosophy, etc) that use those languages. That allows those disciplines to provide a novel understanding of the use of that concept, and hence, once more, a more *exact* understating. Moreover, those disciplines can use it in new applications (e.g., legal AI) or for purely theoretical purposes (e.g., comparing it with concepts of causal justification with different rules of use than the ECtHR's). Hence, the model's rules are more *fruitful* than the concept it models (*fruitfulness*). Moreover, we saw that when we model something, we stupid down reality. That allows us to discard noise and identify the important rules of that concept (important for the use of the model). Hence, we have a simpler but equally effective set of rules of use (*simplicity*). Finally, since the rules of use of the model are a description of the ECtHR's rules of use, then the model uses CAUSAL JUSTIFICATION in ore or less the same way that the Court uses it (*similarity*). Summing up, the criteria of *exactness*, *similarity*, *fruitfulness* and *simplicity* make the model's rules an improved version of the original concept's rules. The foregoing criteria *are* the adequacy criteria of Carnap's *explication* (Carnap 1962).[16]

models as *improved* rules of use

Now that I have made my point as to why modelling ECtHR's causal justifications can be construed as an *explication*, it is time to argue *how to* explicate CAUSAL JUSTIFICATION.

III.3 A recipe of explication: the foundations

To argue how explication should be done, I will use Brun's "*recipe of explication*" as the basis of my methodology (2016, 1226–1229). Brun's "*recipe*" is purposely inexact in order to make room for adjustments depending on the particularities of each explication. This is exactly what I am doing in §3: adjusting Brun's recipe to the

[14] A more complicated model can provide a better justification. This is a simplified example. Still, it does provide an justification independently of its adequacy.

[15] Carnap explicitly acknowledges that *exactness* requires the avoidance of logical inconsistencies so as to achieve better clarity in the use of a concept (Carnap 1963, p.935).

[16] In the literature, some use "*rational reconstruction*" instead of "*explication*", another term that Carnap introduced. However, "*explica-*

particularities of explicating normative judicial justifications with the intention to be used by legal AlgoAI. As a toy example, I explicate the concept of CAUSAL JUSTIFICATION as used in the legal tradition of the ECtHR. In §3.1, I argue which are the source & target system of concepts. In §3.2, I argue how the rules of use should be *improved* based on the four adequacy criteria of SIMILARITY, FRUITFULNESS, EXACTNESS, and SIMPLICITY.[17] Considering that those criteria are epistemic *values*, I proceed by providing specific *checklists* of requirements for each criterion (henceforth *adequacy requirements*). I.e., I apply the proposal of §I.1.2 of providing an *operational definition* of a value in order to perform evaluative judgments for that value. Henceforth, I will use SMALL CAPS to annotate not only concepts, but also the 4 adequacy criteria in order to indicate more clearly that I refer to them. Finally, in §3.3, I argue how explication should start and only how it should start. It is neither realistic nor even recommended to decide *ex ante* every aspect of the practice of explication. What needs to be decided is a first foundation. And that includes *how to begin* an explication.

III.3.1 Source & target systems of concepts

The *source system of concepts* is the system of concepts used in the practice of the judicial authority that the AlgoAI will support or replace (). In our case, that is the ECtHR's system of concepts. As *principal explicandum* I construe the concept that motivates the explication in the first place (in our case, CAUSAL JUSTIFICATION), and as *secondary explicanda* I construe the rest of the concepts from the source system which will also be explicated in order to explicate the principal *explicandum* (e.g., CAUSATION, CAUSE, EFFECT, RESPONSIBILITY, JUSTIFICATION, GENOCIDE, etc.). The respective *explicata* will be called *principal explicatum* and *secondary explicata*.

The *target systems* of concepts is nothing else but the CROSSDI *conceptual interface* of the involved disciplines (§II.2.1, ¶¶7-8). In other words, concepts from the hedgehog-disciplines and the TRANSDI ends are *glued* together to produce new CROSSDI concepts,[18] sometimes by using concepts from the fox-disciplines as glue. The fact that in the target system we glue concepts from different disciplines prompts two challenges related to two distinct types of gluing. The first challenge is that concepts used in one discipline may not be used in other disciplines; there is no concept GENOCIDE in the AI discipline nor the concept DIRECTED GRAPH in legal science. At the same time, such unrelated concepts need to interact in the target system. For instance, if the AlgoAI engineers decide to employ AI modelling methods of CAUSAL JUSTIFICATION that make use of directed graphs like the directed graph of Figure 1,[19] then, for cases of historical negationism like the *Perincek v. Switzerland* case, the nodes of the directed graph should represent information about genocides. To allow such an interaction between DIRECTED GRAPH and GENOCIDE, we need to introduce concepts that will *glue* them. I will name those concepts *gluing concepts*. For instance, to glue the concepts DIRECTED GRAPH and GENOCIDE we can borrow from logic the concept PROPOSITION (gluing concept) that allows us to represent in directed graphs statements about genocides (more on §IV.2).

first type of gluing

The foregoing brings up the subquestion of *how* one should choose gluing concepts. The experts will have to choose candidate gluing concepts and provide a *justification* for their choice. Hence, there needs to be a reference to a *gluing theory* that justifies the gluing. For instance, when one uses PROPOSITION to glue the concepts DIRECTED GRAPH and GENOCIDE, they borrow a concept from formal propositional logic (gluing theory) or from its variations (non-monotonic propositional logic, deontic propositional logic, etc). Note how the gluing concept and theory belongs to the disciplines of the foxes, while the glued concepts (GENOCIDE and DIRECTED GRAPH) to the semantics of the hedgehog-disciplines. Furthermore, since we have a CROSSDI practice, it is the foxes that decide that PROPOSITION is the appropriate glue and the royal hedgehogs that decide what should be the *content* of those propositions using theory from their discipline. At the same time, the AI engineers (the knight hedgehogs) need to make sure that the AI model is compatible with the justifications that are represented by the directed graphs by appealing to theories of their own discipline. Making choices that glue concepts together by appealing to specific theories is a prime example of theory-oriented METADI. Moreover, the practice of gluing concepts is essentially the FRUITFULNESS requirement of *unification*. In brief, I construe as *unification* the gluing of systems of concepts from different disciplinary theories and as *requirement of unification* the requirement of gluing theories together so as to interact FRUITFULLY (more on §3.2.4).

tion" is the more recent term and in contrast with rational reconstruction which is of more general use it is applied *only* in concepts (Leitgeb and Carus 2022, §10. Supplement D: Methodology). The literature that I cite uses "*explication*" and not "*rational reconstruction*".

[17] There are more proposed adequacy criteria in the literature of explication. Such are the criteria of maximising the *explicatum*'s SCOPE OF APPLICATION (Brun 2016, p.1227) and of maximising the number of theories/ontology's/methodologies the *explicatum* UNIFIES (GLUES) (Taylor 2015, o.664 and §7.3). In §3.2.4, I argue that those criteria are in reality requirements of FRUITFULNESS.

[18] Since we have a case of a CROSSDI practice, the TRANSDI ends should be expressed *via* the dominating discipline's system of concepts (*cf.* §I.2.5, ¶9).

[19] For AI modelling methods of CAUSAL JUSTIFICATION that make use of directed graphs *see* Cabalar, Fandinno, and Fink 2014 and its application in Cabalar, Fandinno, and Muñiz 2020 for a *logic-based AI* example; Heinze-Deml, Maathuis, and Meinshausen 2018 for a *probabilistic self-learning* example; O'Shaughnessy et al. 2020 for an *XAI* example.

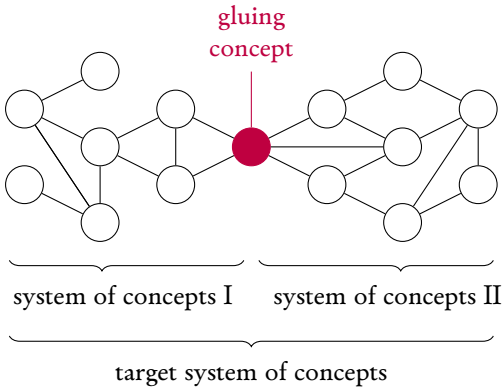


Figure 2: This is a graphical representation of two *glued (aka unified)* systems of concepts. The nodes represent *concepts* and the edges *relations* between concepts.

The second challenge of gluing concepts from different disciplines is that the same term can be used to denote distinct concepts in different disciplines (i.e., the objectivity challenge described in §2.2.1.1, ¶3). E.g., the concept CAUSATION is used both in legal science and in Beckers's 2021 logical causal modelling paper (more on §IV). When this is the case, we need to make sure that whenever the conceptual engineers use that common term (e.g., "*causal justification*"), they all refer to the same concept. In the foregoing example, Beckers's causal model (i.e., a formal explication of CAUSATION) represents rules that are partially different from those of the ECtHR's CAUSATION. Since our goal is to model the ECtHR's causal inference (CROSSDI practice), ECtHR's rules of use *dominate*. In this CROSSDI environment, legal science is the *dominating discipline* and ergo Beckers's rules of use will have to be adjusted according to the explication's adequacy criteria. From the moment that everyone agrees on which concept a term notates, if that concept is part of an already established system of concepts, that system of concepts is glued to the target system (it is *unified* with the target system) and we can use it to articulate new knowledge about the *explicandum*. I will call the concept that glues the two systems of concepts (*see* Figure 2) *gluing concept* as well. second type of gluing

Finally, it can be the case that the term of the *explicandum* changes in the target system. Such is the case for the concept FISH whose term changed from "*fish*" (*explicandum* in the system of concepts of everyday discourse) to "*piscis*" (*explicatum* in zoology's system of concepts) (Brun 2016, p.1220). In CHAPTER IV, we will see that such a change of term should happen to CAUSAL JUSTIFICATION since there are different types of causal justifications suitable for different circumstances which are differentiated by using different terms (e.g., **BUT-FOR** CAUSATION, **NESS** CAUSATION, **ACTUAL** CAUSATION, **INUS** CAUSATION). It should further be noted that in the target system, we can include concepts that reflect the content of the judicial authorities' judgments at a *meta*-level, concepts that are not mentioned by their terms in those judgements. E.g., the Court may use causal justifications to justify its judgements, but it may never use the term "*causal justification*" or similar terms like "*causation*", "*cause*", and "*effect*". Still, from the fact that it uses causal justifications even if it does not acknowledge them explicitly as such, CAUSAL JUSTIFICATION *is* part of the Court's system of concepts. Using such meta-concepts is the common practice of those that analyse legal reasoning: there is no "*non-monotonicity*", "*case-based reasoning*", "*counterfactuals*", "*subsumption*" or "*interpretive concepts*" in the Court's documented judgements (*cf.* §§II.4.1.2-II.4.2; Dworkin 2011, pp.163-164). META-concepts

III.3.2 Improving rules of use

Improving the *explicandum*'s rules of use is the CROSSDI end of making new rules that are more SIMILAR and/or more EXACT and/or more FRUITFUL and/or more SIMPLE. I.e., an improved rule of use is a rule of use that realises more adequately the four *epistemic values* of SIMILARITY, EXACTNESS, FRUITFULNESS, & SIMPLICITY (albeit one can also construe SIMPLICITY as an *aesthetic* value (Baker 2022, §3.3)). Considering this, for the same reasons as for the values of LEGITIMACY, RULE OF LAW, and HUMAN RIGHTS in CHAPTER I, we need to introduce *operational definitions* of the 4 epistemic values. I.e., we need to introduce *checklists* of specific "*factualised*" *adequacy requirements* (henceforth *adequacy requirements* or *explication requirements*). *Specifying* such requirements though is not enough. As Brun 2016 does, we should differentiate between the *specification* of the adequacy requirements and the *evaluation* of whether the model meets those requirements (*ibid.*, pp.1227-1228). The former is part of the *designing & building* phases of engineering and the latter is part of the *analysing* phase (§II.4.1.1, ¶4). Considering that this Thesis is already too long, I will focus mainly on *specifying* adequacy requirements. Fortu- explication/ adequacy requirements

nately, there is a considerable overlap between the ways of specifying and the ways of evaluating the realisation of adequacy requirements. E.g., a way to specify SIMPLICITY is *minimal ontological commitment*: if one has to choose between two *ceteris paribus* adequate *explicata*, they should choose the *explicatum* which makes the fewer ontological commitments. At the same time, this specification condition of SIMPLICITY is an *evaluation* criterion of the *explicatum*'s SIMPLICITY: I evaluate that I made the right choice exactly because that choice requires fewer ontological commitments. Most of the adequacy requirements I introduce are from Brun 2016 (*cf.* Brun 2020). They are in no way exhaustive and they are always open for ever further adjustments, especially during the actual legal ALGOAI engineering practice. Before I introduce adequacy requirements for all four adequacy criteria in §§3.2.1-3.2.4, I would like to make a few general points that hold for all of them.

against mechanisation

The first point I would like to stress is whether there can be *universal* specified adequacy requirements and methods of assessment.. Regarding the *specification* of the adequacy requirements, we should "*make different requirements for different situations.*" (Carnap 1963, p.945). For instance, in legal ALGOAI, the *explicandum* and the *explicatum* should have the *same extension* (SIMILARITY requirement). Otherwise, the AI model does not apply the law in accordance with the ECtHR's judgements and hence we have a violation of legality. However, there are other examples of adequate *explicata* in the disciplines of mathematics, zoology, and philosophy of language for which it is accepted that their extensions are different from the extensions of the respective *explicanda* (Brun 2016, pp.1221-1222; *see also* the adequacy condition of *extensional similarity* in §3.2.1). The fact that there are different requirements *per* case of explication entails that their *assessment* will also more or less differ. Different requirements require different assessment methods. This is one of the reasons why there is no universal *algorithmic* (or *mechanistic* to be *on par* with Enlightenment's maxims in §I.2.3) process for assessment:[20] different methods of assessment would require different algorithms. But even if we narrow down the adequacy requirements to a specific tailor-made set of requirements for a particular case of explication, crafting an algorithmic process of assessment only for the requirements of that set seems *unnecessary*, *unproductive*, and *unattainable*. *Unnecessary*, because as already argued, many times, realising an adequacy requirement is at the same time an assessment of that requirement. *Unproductive*, because as Brun notes, the assessment is a "*creative task*" (*fn.* 30), and hence, mechanising it will impair the said creativity (*cf.* Danks 2014, pp.160-161). For instance, Brun suggests that we should come up with new practical methods of assessment by studying the history of explication and comparing past assessment methods (Brun 2016, *fn.* 30). If everyone was using the same algorithm, then there would not be any different assessment methods to contrast. *Unattainable*, because many times adequacy requirements conflict with each other and hence we have to choose between *explicata* that contribute positively to some of them and negatively to some others. Even if we could determine in advance how potential conflicts can be resolved, and I would do so later on, in general, the resolution of conflicts is *case-sensitive* and hence a full-fledged algorithm would be unattainable. For instance, in the case of a conflict we would have to ask questions like "*How severe is the violation of an adequacy requirement?*" "*Which are the alternatives?*", "*What adequacy criteria do those alternatives corroborate and to what degree?*". Predicting in advance the severity and benefits of every potential violation of an adequacy requirement and then contrasting it with the severity of benefits of violating other adequacy requirements in its place is practically unattainable. Despite all that, I am still open to the possiblity of a (semi-)algorithm method of assessing the realisation of the adequacy requirements, but a method that claims neither universality (i.e., being applicable to every explication) nor perfection (i.e., there is always space for improvements and adjustments to the particularities of each explication), and that is high-level (i.e., its steps are quite general). For more reasons against an algorithmic approach to assessment *see* p.1228.

positive v. negative adequacy requirements

Next, I would like to address the fallacy that for an adequate explication, one should strive to *strengthen* the four adequacy criteria. Let's see a counterexample for the criterion of SIMILARITY. On the one hand, some of the *explicatum*'s and *explicandum*'s aspects must remain similar enough to allow a *comparison* between the two in order to assess that the *explicatum* is indeed an improvement of the *explicandum*. On the other hand though, the *explicatum* must *differ* in certain aspects from the *explicandum*; if they are exactly the same, the former can not be an *improvement* of the latter. And I am not referring only to immaterial differences, but even "*considerable*" ones as we will see in §3.2.1. Hence my choice to keep the word "*similarity*" instead of replacing it with "*sameness*". Considering the foregoing, SIMILARITY requirements can be divided into two categories: (a) *positive* SIMILARITY requirements. I.e., in which aspects the *exlpicandum* and the *explicatum* should be similar; (b) *negative* SIMILARITY requirements. I.e., in which aspects the *exlpicandum* and the *explicatum* should be *dis*similar.

Positive and *negative* adequacy requirements hold for all four adequacy criteria. The main motivation for a negative requirement is a conflict among adequacy criteria. For instance, we will see in §3.2.2 that a negative requirement of EXACTNESS is that in certain cases, the *explicatum*'s rules of use should retain the vagueness, ambiguity, and generality of the respective *explicandum*'s rules of use. What we have here is a *conflict* between the adequacy criterion of similarity (the *explicatum*'s rules of use shall not be different from the *explicandum*'s)

[20] "*Checking whether the explicatum [meets the conditions of adequacy] cannot be done in a* ***mechanical*** *way, but is subject to informal evaluation and judgement*" (Brun 2016, p.1228, emphasis added).

and exactness (the *explicatum*'s rules of use should be more exact). We will see that the criterion of SIMILARITY prevails and hence the *explicatum* should not be more exact than the *explicandum* (negative EXACTNESS). Note that in this example of negative exactness, the negative requirement was for the *explicatum* to not be *more* exact than the *exlpicandum*, i.e., to be *equally* exact. Consequently, by "*negative*" EXACTNESS/SIMPLICITY/FRUITFULNESS/SIMILARITY, I label adequacy requirements according to which the *explicatum* it either *less* or *equally* EXACT/SIMPLE/FRUITFUL/SIMILAR with the *explicandum*.

The foregoing example of a negative exactness requirement indicates that there must be an underlying *hierarchy* among adequacy requirements such that in case of a *conflict* among multiple requirements, the requirements higher in the hierarchy are the ones that win the conflict. Indeed, there exists such a hierarchy. At its top we have the *absolute requirements* and at its bottom the *ceteris paribus* requirements. As *absolute requirements* I construe adequacy requirements that should not be violated in any case unless they conflict with other absolute requirements. Such requirements include the requirements grounded on formal rule of law values (legality, foreseeability, etc) and on minimal substantive legitimacy values like human rights (§§I.2.4,I.2.7). As *ceteris paribus* requirements, I construe requirements that make a difference when choosing the most adequate explication *only if* all other adequacy requirements are equally satisfied by the competing explications. For instance, for Carnap, any simplicity requirement is a *ceteris paribus* requirement since the *explicatum* should be "*as simple as the more important requirements [of exactness, similarity, and fruitfulness] permit*" (Carnap 1962, p.7). *Contra* Carnap, in §3.2.3, I argue that there are cases where SIMPLICITY is not a mere *ceteris paribus* requirement. Quite the opposite actually.

the hierarchy of adequacy requirements

The hierarchy of adequacy requirements is *not* universal and it is *not* totally ordered. The hierarchy is not universal since in different explications we have different hierarchies. For instance, Brun 2016 argues that the resolution of logical inconsistencies or ambiguities seems to be an absolute requirement for *every* explication p.1228. As already argued (*see also* §3.2.2), due to legality, those are *not* absolute requirements. Extensional similarity trumps them. Furthermore, the hierarchy of requirements is *partially ordered*, i.e., two conflicting requirements are on the same level (e.g., two *ceteris paribus* requirements). Finally, considering the above, I propose the following *rule of thumb* to resolve conflicts of adequacy requirements:

RULE OF FAILURE: the degree to which we allow an adequacy requirement to fail is inversely correlated to the strength of the justification that we have for the realisation of the requirement.

After laying out general considerations about all four adequacy criteria, it is time to introduce their operational checklists.

III.3.2.1 Similarity

- *Extensional similarity:* As *extensional similarity* I construe a substantial overlap between the extensions of the *explicandum* and the *explicatum*. Despite seeming like a no-brainer, *extensional similarity* is a rather contestable requirement (*see e.g.* Brun 2016, pp.1220-1222). The classical counterexample is the concept FISH and its explication PISCES in zoology's system of concepts: whales used to belong to the former but they do not belong to the latter since now they are classified as mammals (*ibid.*; Carnap 1962, pp.5-6). This example undermines *extensional equivalence* (i.e., the extensions being *exactly* the same) as an interpretation of extensional similarity. A counterproposal to extensional equivalence has been that the *explicatum*'s extension can differ from the *explicandum*'s extension only if the former includes less objects (Brun 2016, p.1221).[21] This can be motivated by the SIMILARITY requirement of *ontological parsimony*. This counterproposal is still incompatible with actual examples of explication. E.g., 0 was not always considered a number. Ergo, the contemporary conception of NUMBER includes at least one more object, the object 0 (Carnap 1962, p.11).[22] Another counterproposal has been that we should at least concede that there must be a considerable overlap between the two extensions without one being necessarily a subset of the other. Still, actual examples of explication contradict this counterproposal. Brun gives the example of number TWO which has one adequate *explicatum* whose extension is a single object o_1 and another adequate *explicatum* whose extension is another object o_2 such that $o_1 \neq o_2$.[23] I.e., the intersection of the

[21] Brun attributes this counterproposal to Hempel (1988) 2000, p.207 (although Brun uses "*e.g.*" implying that there are more authors that have made a similar suggestion). However, what Hempel says is that the *explicata* "*must not apply to those cases to which the preanalytic concept [explicandum] is* ***generally*** *not applied.*" (¶1, emphasis added). The use of "*generally*" entails that Hempel accepts that it can very well be the case that the *explicatum* applies to cases that the *explicandum* does not apply. This contradicts Brun's interpretation of Hempel.

[22] The reality is even worse. E.g., Pythagoreans did not have a concept of irrational numbers, and hence, the contemporary concept NUMBER has *uncountably infinite* more objects than the Pythagorean one (Kalanov 2013). I am providing this example because Carnap and Brun give counterexamples only of borderline cases of distinct extensions.

[23] For the philopher of mathematics, o_1 is Zermelo's $\{\{\emptyset\}\}$ and o_2 is von Neumann's $\{\emptyset, \{\emptyset\}\}$. For a comparison between the two and

two extensions is *empty* (Brun 2016, pp.1221-1222).

Despite those controversies, due to *legality*, in legal ALGOAI engineering, *extensional equivalence* *is* a SIMILARITY requirement at least for the *principal explicatum*: the extension of CAUSAL JUSTIFICATION should not change after its re-engineering. More precisely, we saw that unless granted permission by judicial authorities to make a change, the ALGOAI model has to be a *descriptive* model of the ECtHR's practice (§II.4.1.2). Any violation of this requirement is a violation of legality (§I.3.2.1.2, ¶14). Consequently, in order to be a description of how the ECtHR interprets & applies the law, the extension of CAUSAL JUSTIFICATION needs to be exactly the same in both source & target systems of concepts. Furthermore, due to legality being the most fundamental rule of law requirements (§I.2.4, ¶1; §I.2.3, ¶9), *extensional equivalence* is an *absolute* explication requirement: in case of conflict with other explication requirements it always prevails. Note that Brun *does* acknowledge that extensional equivalence can be a requirement for explication. An "*extreme*" one that "*[leads] to limiting cases of explication*", but still a valid SIMILARITY requirement (Brun 2016, p.1222).

- *Intentional similarity:* An *intentional* definition of a concept is a definition that specifies *necessary & sufficient* conditions for correct application of the concept (Cook 2009, p.155). In the context of the logical model of interpreting and applying the law in §II.4.1.2, intentional similarity can be construed as the *explicandum* and the *explicatum* having similar necessary & sufficient conditions for their subsumptive tests. Those conditions can be transformed into IF-THEN rules of use like in the case of logic-based expert systems (§II.4.2.1; *cf.* §I.3.2.1, *fn.* 79).[24] The translation to an IF-THEN rule of a *sufficient* condition $\mathcal{I}$ can have the following logical form: "*IF condition $\mathcal{I}$ holds for a particular t, THEN t is subsumed by the concept $\mathcal{C}$.*". The translation to a IF-THEN rule of an *necessary* condition $\mathcal{I}$ can have the following logical form: "*IF condition $\mathcal{I}$ does not hold for a particular t, THEN t is not subsumed by the concept $\mathcal{C}$.*". In the *intentional* rules of use I also include rules that dictate when a condition is *not* sufficient/necessary for a particular t to be subsumed by a concept $\mathcal{C}$. They can be coined as *negative intentional* rules of use *contra* the *positive intentional* rules of use.

  more translations to rules

  Changes in the intention of a concept can lead to changes in the extension of a concept. For instance, if we consider that the condition of a causal relation being asymmetric like in Figure 1 is not necessary, the extension of CAUSAL RELATION can be expanded to include *both* symmetric and asymmetric relations. Consequently, in order to satisfy the absolute requirement of *extensional equivalence*, we need to make sure that the *explanatum*'s intention is similar enough to the *explanandum*'s intention so as to share the same extension.

- *Similar formal form:* Brun recommends to specify the *logical form* of the *explicandum* and of other concepts that will be explicated (Brun 2016, pp.1226-1227). As an example, they provide the logical form of the concept HARDER which is used to explicate the concept MINERALOGICAL HARDNESS as it is expressed in Schumann 2008, p.20 (remember also the different logical forms of *ought*- and *is*-statements that spawned the subdiscipline of deontic logic (§I.2.3, ¶7)). Its logical form is a two-ary predicate that compares two objects (*comparative concept*):`harder(x,y)` (Brun 2016, p.1229).[25,26] Regarding the different types of logical forms, Carnap has argues that a concept's logical form is either a function or a predicate or a relation, all of which are common means of representing concepts in formal logic (Carnap 1962, p.8).

  In §1.1 and *fn.* 4, I argued that ALGOAI engineers should use more formal languages than logic, semi-formal ones to be precise. Therefore, what they need to specify is the *(semi-)formal form* of those concepts which does not have to be logical; it can be any adequate expression of the *explicatum* formed by the (semi-)formal language $\mathcal{L}$ they use. Also, note that Brun seems[27] Finally, I consider the introduction of a form to be a matter of *syntactical interpretation* (*cf.* Prakken 1993, p.14).

---

an account that compromises both positions of TWO's extension have a look at the Benacerraf's 1965 seminal that we saw in §I.1.1, ¶5 & §I.1.1.1. The common citation in all those references is no accident. They all refer to problems of ambiguity of meaning like what is the extension of TWO. I.e., they all refer to instantiations of the *objectivity challenge*.

[24] Remember that this translation is not always possible (§II.4.1.2, ¶4).

[25] Typically, predicates are *non*-logical symbols. I assume then, that by "*logical form*', Brun means a *well-formed* formula built with both logical symbols and non-logical symbols from a signature.

[26] One could argue that since Brun has construed concepts as a term t and a set of rules $\mathcal{R}$, the logical form of a concept should be a two-ary function $\mathcal{C}$ whose arguments are the term and the set of rules: $\mathcal{C}(t, \mathcal{R})$. Always by being lenient as to what constitutes a *logical* form since accepting the set $\mathcal{R}$ in our ontology entails that we have incorporated set theory's language in our "*logical*" form. Both alternatives are correct. This is another example of how the same concept can be expressed differently depending on the language $\mathcal{L}$ we are using (*cf.* §II.4.1.1).

[27] I write "*seems*" since Brun does not say so explicitly. What Brun does is that instead of arguing that the predicate `harder/2` is ***one of the possible*** logical forms of the concept MINERALOGICAL HARDNESS, they argue like `harder/2` is the ***only*** proper logical form of MINERALOGICAL HARDNESS. to talk about a *unique* logical form something that I reject since: (α) it precludes the possibility of finding more adequate forms; (β) the same rules of use can be expressed *via* different (semi-)formal languages $\mathcal{L}_i$ (*cf. fn.* 26; §II.4.1.1).

- *Relational similarity:* This requirement is about *identifying* concepts in the source system with which the *explicandum* is *related*, as well as the *nature* of those relations. Take the example of CAUSAL JUSTIFICATION. In the source system, the concept CAUSAL JUSTIFICATION is related to the concept JUSTIFICATION: the former is a *hyponym* of the latter. This entails that the rules of use of "*justification*" are also rules of use of "*causal justification*" and that the latter has some extra rules that determine its *causal* character and differentiate it from other non-causal justifications. Considering this, a tactic to perform an explication is to identify hypernyms of the *explicandum* for which we already know some rules of use, and then, try to figure out which are the extra rules that differentiate the *explicandum* from its hypernym. To identify which are those extra rules of use, we can compare the *explicandum* to other hyponyms of the same hypernym to understand why they differ (e.g., comparing which rules differentiate CAUSAL JUSTIFICATION with DEFINITION and GROUNDING).[28]

- *Coherence:* There are multiple ways to construe coherence in the context of judicial reasoning. For an example of the different approaches of coherence in the ECtHR's legal tradition *see* Letwin 2021. To be *on par* with the rest of the Thesis, I will construe coherence as the decision of which *subsumption criteria* we should choose for a concept C so as to satisfy the conjunction of multiple *paradigmatic cases* of C's application (*cf.* §II.4.1.2). E.g., identifying subsumption criteria about the concept CAUSAL JUSTIFICATION in order for the concept to subsume paradigmatic cases of causal justification in the ECtHR's case-law. In the literature of judicial reasoning, this is a specific type of the legal interpretation method called *narrow reflective equilibrium* (Schroeter, Schroeter, and Toh 2020, p.94; *see also* the introduction to reflective equilibrium in Daniels 2020). RE is another method of conceptual engineering which according to Brun 2020 complements explication making up for its deficits. Brun 2020, at the end of their paper, advocate for further research on how RE and explication can be merged. In §3.3 &§IV.3, I provide a step towards this direction. But first things first, what *is* reflective equilibrium?

  Reflective equilibrium (*RE*) is the method of *reflecting* on the already available paradigmatic knowledge so as to find the appropriate *equilibrium* among competing justifications of that knowledge (Daniels 2020, §1). *Narrow* RE (*contra wide* RE) is when we reflect on the coherence of paradigmatic knowledge in a particular system of concepts (in our case the ECtHR's legal tradition of human rights law) and we do not compare this coherence to the coherence induced by alternative legal/moral/philosophical theories (e.g., to the utilitarian Anglo-American jurisprudence) (*ibid.*, §3). RE originates from Goodman's seminal 1955 "*Fact, fiction, and forecast*" book on the justification of inductive logical inferences (*ibid.*, §2.1; *cf.* Brun 2020) and the term "*reflective equilibrium*" was coined by Rawls in his landmark "*A theory of justice*" (1971) (Daniels 2020, §2.2). Rawls' conceptual re-engineering of the concepts JUSTICE and RIGHTNESS (*fn.* 6) is performed *via* RE.

  To wrap up, in the context of explication, the *coherence* requirement can be construed as the identification of rules of use that satisfy multiple paradigmatic cases of the concept's application. The subset of those rules which are necessary and/or sufficient for the satisfaction of the paradigmatic cases can be construed as a subset of a concept's *intentional* rules of use. I.e., *coherence* can be used to decide which *intentional* rule of use we will explicate, and subsequently, *coherence* provides a *justification* of why a rule of use is an *intentional* rule of use. As we will see in §3.3, this can be quite useful when we want to demarcate between genuine *v.* non-genuine *intentional* rules of use.

III.3.2.2 Exactness

- *The threat (?) of logical paradoxes:* According to Carnap, a requirement for the adequacy condition of exactness is the resolution of any logical paradoxes emerging from the *explicandum*'s rules of use in the source system of concepts (Carnap 1963, p.935). *Prima facie* it seems indeed that logical paradoxes should be resolved since they undermine the legitimacy value of *legal certainty*: how can we be certain about how the law is applied if the law is paradoxical (*cf.* §I.2.4, ¶5)? Having said that, *contra* Carnap's position, in certain legal ALGOAI engineering cases, those paradoxes should be retained. As we will see, instead of fixing them, ALGOAI engineers should *embrace* them. Note though, that even if a logical paradox is malignant and it should not be accommodated in legal ALGOAI, this is a decision that should be decided by the judicial authorities and *not* by the ALGOAI engineers. Otherwise, we have an *illegitimate* exercise of judicial power (§§I.3.2.1.2, ¶13-I.3.3).

  What is a *logical paradox* though? Let's start with the definition of a *paradox*. It is an *argument* for which

[28] Brun briefly refers to the idea of using *hyponyms* to perform an explication but without any further elaboration for the *why* & *how* like I do.

the following conditions hold (Cook 2009, p.214):

(C_1) The argument has "*apparently*" true premises.

(C_2) The argument's conclusion is reached *via* an "*apparently*" unobjectionable reasoning

(C_3) The argument's conclusion is false or even contradictory.

Condition (C_3) is about the conclusion being either contradictory or false. Based on that, I construe as a *loigcal paradox* the paradox whose conclusion is a *logical contradition*, where a *logical contradiction* is when the conclusion of an argument is both true and false at the same time. In other words, for a paradox to be a logical paradox, condition (C_3) is specified in the following form:

(C_3)′ The argument's conclusion is *logically* contradictory.

Considering the above, the source of the paradox's paradoxicality is that either the "*apparently*" true premisses are not true after all (i.e., condition C_1 does not hold) or that the "*apparently*" unobjectionable reasoning is not unobjectionable after all (i.e., condition C_2 does not hold) or both of the foregoing. This distinction allows us to classify logical paradoxes into two categories: (a) *conceptual* and hence *semantical* paradoxes; (b) *syntactical* paradoxes.[29] *Semantical* logical paradoxes are the logical paradoxes whose paradoxicality is grounded on the falsehood that the premisses are true. *Syntactical* logical paradoxes are the paradoxes whose paradoxicality is grounded on the falsehood that the reasoning *via* which the conclusion is reached is unobjectionable. In what follows, I provide some common cases of semantical and syntactical paradoxes in judicial reaosning using examples from the ECtHR's case-law.

SYNTACTICAL PARADOXES: In *syntactical paradoxes*, the premisses of the paradoxical argument are true. However, the reasoning *via* in which the Court reaches the logically contradictory *conclusion* is objectionable. I.e., it is a problem of the argument's *logical syntax*, where as logical syntax I construe *both* the syntax used to structure the premisses and the conclusion of the argument, as well as the inference rules used to derive the conclusion from the premisses. Since the Court does not derive judgements mechanistically, we should expect the Court to neither make the logical relations in its arguments explicit and precise nor use the exact same phrasing for the arguments that it repeats. Thus, in order to identify logical contradictions induced by the syntax of arguments, we have to resolve to a *syntactical* interpretation of those arguments. I.e., we have to interpret the syntax of the Court's arguments which is many times implicit, inexact, and inconsistent.[30]

The definition of syntactical paradoxes will become clearer by looking at two typical types of syntactical paradoxes in legal reasoning, the paradoxes that I will call *counterfactual* paradoxes *&* *monotonic* paradoxes:

I. *Counterfactual paradox:* According to the rule of law's *principle of equality*, similar cases should be treated similarly and dissimilar cases should be treated dissimilarly (CDL-AD(2016)007, ¶70; *cf.* §I.2.4, ¶2; §II.4.2.1, ¶4). There is a lot of debate in the literature as to what would make two cases similar so as to require the same treatment (*see e.g.* Walton 2002, §1.10). A *minimal* requirement is that from *the same premisses the same conclusion should follow*. I construe as *counterfactual paradox* the paradox where in different cases from the same true premisses contradictory conclusions follow. For instance, Mchangama and Alkiviadou argue that while in the cases of Holocaust denial it is usually implicitly assumed that the denial of the Jewish genocide (premiss) entails incitement to hatred or intolerance (conclusion), in the *Perincek v. Switzerland* case, the Court judged explicitly that the denial of the Armenian genocide (premiss) does *not* entail incitement to hatred or intolerance (conclusion) (2021, p.1022). Ergo, from the same premiss ("*The applicant denied a genocide.*"), the Court reached two different logically inconsistent conclusions.

counterfactuals biases

Counterfactual inconsistencies can be used to identify biases in the Court's judgements (remember also the use of counterfactual inconsistencies in order to engineer *fair* ALGOAI models in §I.2.5, ¶8; *cf.* §II.4.2.2). More precisely if the same counterfactual inconsistencies appear in multiple cases which have common characteristics $\mathcal{X}_i$ (e.g., te applicants have the same gender, race, country of orgigin, etc), then one can raise the question of whether it is in virtue of those characteristics that that judicial authorities judge inconsistently. I.e., if for a specific value $\mathcal{X} = x$ of a characteristics $\mathcal{X}$ the Court derives $\mathcal{E}$ from the set of premisses $\mathcal{C}$, while for the *counterfactual* case in which $\mathcal{X} = x'$ with $x \neq x'$ it derives $\neg\mathcal{E}$ from the same premisses $\mathcal{C}$, then it may be that the Court makes an irrational inference due to bias with regards to $\mathcal{X}$.

judicial authorities trumping logicians

[29] I am following the contrual of *semantical* and *syntactical* interpretation from §2.2.1.1, ¶6.

[30] "*Although the judge's expressed reasoning is informal and to a degree enthymematic, it is quite capable of being re-cast without loss or gain in a more rigorously and formalistically [logical] form.*" (MacCormick 1992, p.184).

The question then becomes which of two conclusions is true: $\mathcal{E}$ or $\neg\mathcal{E}$? One more, this is the job of the judicial authorities to decide, not of the AlgoAI engineers.

For instance, Mchangama and Alkiviadou 2021 (pp.1022-1023) argue that the Court's judgements are more strict when it comes to Holocaust denial cases in comparison with other cases of historical negationism like the denial of the Armenian genocide. This difference was also highlighted in the dissenting opinion of judges Vuinić and Pinto de Albuquerque in *Perincek v. Switzerland (2013)*, ¶22: "*[t]he suffering of an Armenian under the genocidal Ottoman State policy is not worth less than the suffering of a Jewish person under the genocidal Nazi State policy.*". Ergo, one can argue that the Court is biased by being more protective towards the rights of the victims of the Holocaust *contra* the rights of the victims of other genocides. The decision of which right should be restricted so as to solve the paradox, the right to freedom of expression of those denying genocide or the right to dignity of the victims of the genocide, is *not* a decision that an AlgoAI engineer should make.

II. *Monotonic paradox:* Classical logic has the so-called property of *monotonicity* according to which if from a set of premisses C one can draw a conclusion $\mathcal{E}$, then for every set of premisses which includes C they can derive the same conclusion $\mathcal{E}$. One would expect the same for legal inferences, albeit that is not the case. For instance, Mchangama and Alkiviadou argue that while the denial of Armenian genocide in the *Perincek v. Switzerland (2015)* case is similar to the Holocaust denial in other ECtHR cases, in the *Perincek v. Switzerland (2015)* case the Court used *additional* contextual factors (e.g., the historical relevance of Armenian genocide to Switzerland) that lead to a different judgments than those in similar Holocaust denial cases. In other words, by adding *additional premisses* to the arguments used in the Holocaust denial cases, the conclusion of the Court changed (Mchangama and Alkiviadou 2021, pp.1022-1023).

Despite Mchangama and Alkiviadou's objection, the use of non-monotonicty does not raise *per se* concerns about violations of the principle of equality or of unjust reasoning in general. Non-monotonicity is usually construed as an inherent benign aspect of legal inference without of course excluding the possibility that indeed there are cases in which it is misused/abused (*see e.g.* Poggi 2021, pp.428-429), Such a benign example of monotonic paradoxes is arguments whose conclusion is *pro tanto*. I.e., it "*may be withdrawn on the basis of further information*" (Sartor 2012, p.112). Such *pro tanto* cases are cases where a court *lacks knowledge* (Gordon 1988) what Poggi 2021 (p.427) calls *epistemic deficit*. For instance, based on the principle of the presumption of innocence, unless there is sufficient evidence, the defendant should not be judged guilty. However, in the face of additional evidence, i.e., additional premisses, a court can deem the defendant guilty.[31] Other cases of accommodating non-monotonic reasoning quite widespread in legal practice are: (α) adding *exceptions* to a previous conclusion (p.427-428; *see also* Gordon 1988, pp.113-116; Rigoni 2014, pp.37-48).[32]; (β) resolving *disputes* of competing arguments where new premisses defeat the previous argument (Gordon 1988, p.113; *cf.* Governatori, Rotolo, and Sartor 2021, p.691; §I.2.4, ¶3).

Semantical paradoxes: Regarding *semantical* logical paradoxes, we saw that their paradoxicality is grounded on the fact that some of their *premisses* are not true. I.e., it is an issue of *semantical interpretation*. Following the logical explication of the *interpreting & applying* of the law in §II.4.1.2, I will focus on *conceptual* semantical paradoxes. I.e., paradoxes that concern the *subsumptive tests* which determine which particulars are subsumed by which concepts. Ergo, those paradoxes are essentially faulty (?) *interpretations* of the law which lead to faulty (?) *applications* of the law. Three common types of conceptual interpretation challenges in the literature are: (α) *vagueness*; (β) *ambiguity*; (γ) *generality*. In what follows, I elaborate on the particularities of semantical paradoxes induced by the foregoing typology using Fine's 1975 disambiguation of that typology (*ibid.*, pp.265-267; *compare* with the following citations on vagueness, ambiguity, and generality in legal reasoning which are not *per se on par* with Fine's citation: Sorensen 2022, §2; Marmor 2018; Poscher 2012; Walton 2002, pp.69-72; Schauer 1991, §2.7).

I. *Vagueness:* some of the premisses are neither true nor false Those premisses are characterised as *under*-determined and this phenomenon is characterised as *vagueness*. More precisely, it is under-determined

[31] Poggi argues that the *pro tanto* approach is not compatible with legal reasoning since judges do not "*[draw] conclusions tentatively, reserving the right to retract them*" (Poggi 2021, p.429). In their argument, Poggi wrongfully interprets a *pro tanto* conclusion as a *tentative* conclusion. That can indeed be one interpretation, but one among many. A court can *non*-tentatively reach a conclusion, and then, *non*-tentatively re-evaluate this conclusion in future cases. E.g., when a court convicts a criminal, they do so non-tentatively. However, it may be the case that in the light of new evidence or in case of appeal to a higher court, a new court finds the defendant not guilty. The case of the ECtHR is also a case of non-tentative *pro tanto* judgements due to the living instrument interpretation dogma (§II.4.1.2). Specifically, for a certain period of time, the Court non-tentatively judges according to the then present-day standards of the HCPs. At a later point though, if those standards do change, it will once more non-tentatively re-evaluate them.

[32] Poggi 2021 (p.327, footnote 40) provides a list of citations with arguments against the necessity of using non-monotonicity to accommodate for exceptions in legal reasoning.

whether a particular fact is subsumed by a concept of a legal norm. For instance, ARTICLE 14 (PROHIBITION OF DISCRIMINATION) prohibits discrimination grounded on "*sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or* ***other status.***" (emphasis added). What counts as "*other status*" is *under*-determined. E.g., it is not specified in the law that the particular "*sexual orientation*" is subsumed by the concept OTHER STATUS. Consequently, the truth value of the proposition "*ARTICLE 14 protects against discrimination of sexual orientation.*" is under-determined and it is up to the Court to interpret it as true or false based don't the living instrument doctrine. And the Court interpreted it as true (ECtHR's Press Unit 2023; *cf.* Byron 2016; Jung Lee 2022).

II. *Ambiguity:* some of the premisses take multiple truth values at the same time. Those premisses are characterised as *over*-determined and this phenomenon is called *ambiguity*. More precisely, it is both true and false that a particular is subsumed by a concept and hence the subsumption test is overdetermined (or as Poscher 2012 (p.129) puts it, the respective concept has more than one meanings). For instance, in the *Perincek v. Switzerland (2015)* case, the Court argued that the number of the addresses of an expression of historical negationism is correlated to the harm induced by that expression: fewer audience entails less harm (¶254). At the same time, in the *Witzsch v. Germany (2005)* case, the Court argued that the fact the expression of historical was "*made in a private letter and not before a larger audience is* ***irrelevant***" (¶4, emphasis added). Thus, in the *Perincek v. Switzerland (2015)* case, the small number of audience contributes negatively to the subsumption of a particular expression under the concept VIOLATION OF THE CONVENTION, while in the *Witzsch v. Germany (2005)* case, it does not have any influence to the decision of its subsumption. Hence, the proposition "*The fewer the audience the lesser the harm.*" is true in the first case and false in the latter. Note though that not every case of conflicting norms is undesirable since as as we saw in §I.2.4, ¶3 *defeasible reaosning* is an inherent component of judicial reaosning.

III. *Generality:* This is the case where interpretation of a concept is too *general* ending up including more particulars than what it should. For instance, ARTICLE 10 protects freedom of *expression*. In the process of deciding whether there has been a violation of ARTICLE 10, the ECtHR needs to identify whether the act of the applicant that caused the alleged harm is indeed an *expression* (e.g., an expression of historical negationism, an expression of military information, an expression of whistle-blowing) (ECtHR Registry 2021). However, the Court has considered many times as particulars of EXPRESSION cases of ***physical conduct*** (*ibid.*, ¶14). For instance, in the *Shvydka v. Ukraine (2014)* case, the applicant detached part of a ribbon bearing the phrase "*the President of Ukraine V.F.Yanukovych*". According to the Court, this conduct "*meant to* ***express*** *her opinion that Mr Yanukovych could not be called the President of Ukraine*" and hence the Court judged that its protection under ARTICLE 10 is admissible (*ibid.*, ¶8, emphasis added). The question as to whether *physical conduct* can be subsumed by EXPRESSION and hence whether it should be protected under freedom of expression provisions is a rather controversial topic since doing so leaves a big room for misuse of power considering that in principle "*speech enjoys greater immunity than conduct*" (Barendt 2019, §5). Ergo, acts of conduct that should be restricted are in turn legally permissible.

Considering the above examples of semantical and syntactical paradoxes, it should be highlighted that ALGOAI engineers should strive to *identify* them, but they should not always strive to *resolve* them. Firstly, as argued in §I.3.2.1.2, ¶14, they should refer those paradoxes to the human judicial authorities who should have the final say about how to resolve them. ALGOAI engineers have the obligation to retain them unless judged otherwise by the judicial authorities, something that entails that the requirements of *extensional equivalence* (§3.2.1) trumps the criterion of EXACTNESS. Secondly, many of those paradoxes are actually *desirable* in judicial practice. For instance, I provided examples of vagueness and non-monotonicity (*fn.* 31) that allow the ECtHR to implement the interpretation dogma of the *living instrument*.

- *Explicitness v. implicitness:* Carnap recommends making *explicit* the *implicit* rules of use of the *explicandum* (Carnap 1962, p.7). Having said that, making explicit implicit knowledge can have a considerable impact on the *complexity* of the model (ergo, a conflict with the SIMPLICITY requirement of *minimising complexity*). Take the example of (legal) expert systems during the second AI spring. The engineers that were engineering an expert system for a specific discipline (e.g., law) had to turn implicit knowledge used by the disciplinary experts to a multiplicity of IF-THEN rules ending up engineering models with prohibitive complexity, a problem that had a decisive impact on the emergence of the second AI winter (Duchessi and O'Keefe 1995; Vedder, van Dyke, and Prybutok 2002; *cf.* §I.3.2.1, *fn.*79). Consequently, there needs to be a balance between which rules are made explicit and which remain implicit. It is also recommended to make explicit during the designing phase implicit rules that will not be used eventually in the model, and then, during the building phase, to choose which of those will be disregarded so as to build the "*stupid*"

model (*cf.* the SIMPLICITY requirement of *representational stupidity*). The choice of which rules will become explicit can be done by using the adequacy requirements like making explicit the rules of use that satisfy adequately the SIMILARITY requirement of *coherence*.

III.3.2.3 Simplicity

- *Syntactical simplicity:* This requirement is about using *simpler* expressions in the semi-formal language $\mathcal{L}_d$. This is the type of simplicity that Carnap had in mind when he introduced the four adequacy criteria (Brun 2016, p.1224). For instance, Carnap proposes to evaluate the syntactical simplicity of the *explicatum*'s *formal form* or of the *relations* that it has with other concepts in the target system (Carnap 1962, p.7).

- *Semantical simplicity (or representational stupidity):* We saw that some of the model's constitutive elements need to be semantical representations of reality, but of a *stupid down* version of reality. We also saw that the decision of the level of how stupid a model should be guided by the model's TRANSDI ends (§1, ¶4). The appropriate level of stupidity contributes positively to *understanding* the elements of reality that are actually relevant to the TRANSDI ends, to *communicating* the results to non-experts, as well as to make disciplinary knowledge (e.g., semantics, methodologies, theories) more epistemically accessible to experts from different disciplines. These positive contributions corroborate the FRUITFULNESS requirements of *maximizing understandability & explanatory power*. We also saw that the level of stupidity should be balanced according to the *computational needs* of the intended application (§1, ¶5). This corroborates the SIMPLICITY requirement of *minimising complexity*.

  Note that simplifying reality can be achieved by more strategies than merely cutting down the elements of reality that are represented by the model. For instance, it can also be the case that multiple aspects of reality are represented by a single part or relation of the model. For instance, assume that a forest has many dry trees and flowers. We can represent all of them by the term "*dry vegetation*" as done in Figure 1. Ergo, representational stupidity is also correlated to the FRUITFULNESS requirements of *maximizing universality & informational power*.

- *Parsimony requirements:* The gist of the parsimony requirements is that the ALGOAI engineers should make as few *commitments* (e.g., *ontological* commitments) as possible; the more commitments the more restriction to the model's FRUITFULNESS requirements of *maximizing universality & understanding*, as well as to the SIMPLICITY requirement of *minimizing complexity*.

  Based on the typology of METADI practice in §II.3.1.2.1, we can extrapolate the following three types of parsimony: (α) *Metaphysical parsimony:* It is about staying as metaphysically *neutral* as possible. For instance, in §2.2.1, *fn.* 8, I construed CONCEPT in a way that avoids certain commitments about *what* concepts *are* (e.g., abstract entities, mental representation). And by doing so, I made my approach compatible with more metaphysical theories than what would have been the case otherwise. A special type of metaphysical parsimony is the already introduced *ontological parsimony:* we should not burden our ontology with more objects than what is necessary to satisfy the rest of the adequacy requirements.; (β) *Theoretical parsimony:* I reduce theoretical parsimony to at least the following three requirements: minimizing the terminological load by choosing as fewer theoretical terms as possible, minimizing the number of theories invoked from the different disciplines of the target system, and choosing the less complex theories to make our case. (γ) *Methodological parsimony:* I reduce methodological parsimony to at least the following two requirements: minimizing the number of methodologies employed in the METADI level as well as choosing the less complex methodologies (for *complexity see* the next SIMILARITY requirement).

  I would also like to add another type of parsimony in juxtaposition to metaphsyical parsimony: *(δ) epistemic parsimony*. I construe epistemic parsimony as a twofold requirement: ($\delta.i$) *expert*-oriented epistemic parsimony: minimizing the knowledge we expect the disciplinary experts that explicate a concept to have about other disciplines. That way we can expand the range of experts that can do the job.; ($\delta.ii$) *user*-oriented epistemic parsimony: minimizing the knowledge we expect the *users* of the *explicatum* to have about the disciplines involved in the explication of the *explicandum*. E.g., if a formal model of CAUSAL JUSTIFICATION is used by judges, we should not expect judges to know much about logic or AI engineering. *See also* the FRUITFULNESS REQUIREMENT of *maximising understandability*.

  Concluding, note that the different types of *parsimony* overlap. E.g., a theory carries its own set of metaphysical, methodological, and epistemic commitments.

- *Minimizing complexity:* With *complexity*, I refer to the classical account of time complexity in computer science: between two ways of realising a task, i.e., between two algorithms, the algorithm with the more

steps is the more complex one (Dean 2021, §1.2). Algorithms with higher complexity will produce an output at a later time (or sometimes they will never produce an output since they will never halt) contradicting the legitimacy requirement of delivering judgments in a *timely* manner (§I.2.4, ¶8).

There is a misconception that SIMILARITY requirements are *ceteris paribus* requirements (*see* e.g. Carnap 1962, p.7; Brun 2016, p.1224; *contra* Goodman 1958). I have already argued how SIMILARITY requirements corroborate (or undermine) other adequacy requirements (e.g., the FRUITFULNESS requirements of *maximising explanatory power, informational power, universality, understanding*). Ergo, they are not *ceteris paribus* requirements; it is not the case that "*all other adequacy requirements are equally satisfied by the competing explications*" (§3.2, ¶6).

III.3.2.4 FRUITFULNESS

There are two characteristics of FRUITFULNESS that distinguish it from the rest three adequacy criteria. The first one, is that FRUITFULNESS is a heavily *contextual* adequacy criterion. The same *explicatum* can be very fruitful in one setting and very *un*fruitful in another. For instance, an explication of CAUSAL JUSTIFICATION written in heavy-loaded legal terminology may be very informative for lawyers that want to use it in their practice, but it can be confusing and unintelligible for the AI engineers that want to formalise it in AI models. For the rest of the adequacy criteria (SIMILARITY, EXACTNESS, SIMPLICITY), it is usually the case that they apply in all settings. E.g., a logically consistent syntactically simple *explanatum* that is extensionally equivalent to the *explanandum* will continue being logically consistent, syntactically simple, and extensionally equivalent no matter *who* uses it and *for which* purpose. This is because SIMILARITY, EXACTNESS, SIMPLICITY are more or less *user*-independent and hence *practice*-independent. FRUITFULNESS requirements though are very much *practice*- and *user-dependent*. The same rules of use are more or less fruitful depending on the practice and even on the individual practitioners of that practice (e.g., from the same explication of CAUSAL JUSTIFICATION written in heavy-loaded legal terminology some judges will extrapolate more information than others).

FRUITFULNESS: user- & practice-dependent

The second characteristic that distinguishes FRUITFULNESS from the rest of the adequacy criteria is that it is the adequacy criterion that is more strongly correlated to *use* of the model: a fruitful model is a *useful* model. Let's see why this is the case in more detail:

- *Realising the TRANSDI ends:* In our case, this requirement is about realising the checklist of LEGITIMACY's operational definition. Those TRANSDI ends will *inter alia* guide us to decide which adequacy requirement we should choose in case of conflicts among requirements. E.g., we saw that in many cases, resolving logical paradoxes is necessitated from the TRANSDI requirement of *foreseeability*, and hence, in case of conflicts with the requirement of syntactical simplicity, the resolution of a logical paradox should prevail; we should choose the more syntactically complex option since it corroborates foreseeablity. Note once more that the introduction of the TRANSDI ends in the target system should be done *via* the source system of the dominating discipline (*fn.* 18).

- *Maximising universality v. maximising informational power:* For Carnap, a requirement of FRUITFULNESS is the ability to formulate *universal statements* (Carnap 1962, p.7). Or in non-philosophical *lingo*, the ability to *generalise* (e.g., Brun 2016, p.1227; compare the requirement of *universality* with the EXACTNESS requirements of resolving paradoxes of *generality* in §3.2.2). For instance, we saw that Figure 1 is applicable to every instance of lightning hitting a tree in a forest with dry vegetation (§2.2.2). The model is not bounded by the particularities of a specific cases (e.g., the spatiotemporal characteristics of a specific forest fire), but it is *generalisable* to multiple such cases. Such a model belongs to the kind of universal statement that Carnap characterises as *empirical laws* (*ibid.*).

  Considering the above, the *universality* of the *explicatum* will be the ability of the rules of use to be applicable to multiple cases. The cases in which a rule of use is applicable will be the *scope of application* of that rule. The wider the *scope of application* the greater the *universality* of the respective concept.[33] However, if it becomes too wide, it may not be of practical importance regarding the realisation of the TRANSDI ends (FRUITFULNESS requirements). Look for instance the two models of Figure 3:

  The two models are ordered from left to right in terms of *universality*. However, in terms of the amount of *information* they provide to the user (henceforth *informational power*) the order is reversed. More precisely, both models constitute toy justifications where the tail of the arrow *explains* why the head

[33] Note that Brun considers the *scope of application* as an extra adequacy criterion next to FRUITFULNESS, EXACTNESS, etc. At the same time, they construe FRUITFULNESS as *universality* (Brun 2016, p.1227). Personally, I can not see how FRUITFULNESS as *universality* and *scope of application* can be separated. A scope of application with a wide range is by definition what gives a universal statement its universality.

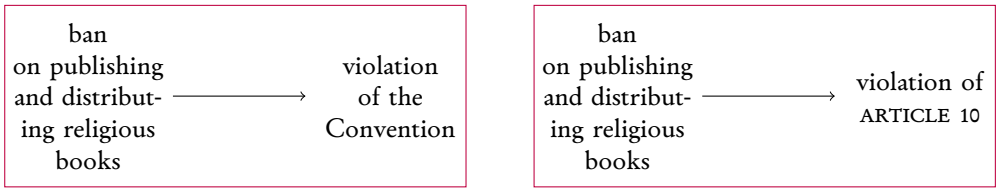


Figure 3: Two models where the *tail* of the arrow *explains* the *head*. MODEL I is more *universal* (or more *general* if you prefer) than MODEL II, while MODEL II provides more *information* than MODEL I. The example is taken from the *Ibragim Ibragimov and others v. Russia (2018)* case (*see also Taganrog LRO and others v. Russia (2022)*).

of the arrow is the case. The tail is the same in both models: a state-actor banned the publication and distribution of religious books. The head though differs: in MODEL I, the head shows that the banning constitutes a violation of the Convention, while in MODEL II, the head shows that the banning constitutes a violation of a *specific* article of the Convention, ARTICLE 10 (FREEDOM OF EXPRESSION). MODEL I can be applied to multiple articles of the Convention, not *per se* to ARTICLE 10. E.g., banning the publication and distribution of religious books can under certain circumstances constitute a violation of ARTICLE 9 (FREEDOM OF THOUGHT, CONSCIENCE AND RELIGION) (*see Ibragim Ibragimov and Others v. Russia (2018)*). Ergo, MODEL I has a *wider* scope of application than MODEL II and hence its *universality* is higher. At the same time though, knowing that the Convention was violated does not entail that it was ARTICLE 10 that was violated, while knowing that ARTICLE 10 was violated *does* entail that the Convention was *also* violated. In other words, MODE II provides the same *and more* information than MODEL I and hence it has higher *informational power*. Once more, the choice of optimal balance between maximising universality and maximising informational power will be decided based on the TRANSDI intended application: what type of information do we want the model to provide and in which cases we want to be applicable (i.e., which should be its scope of application)?

overfitting *v.* underfitting

At this point, I would like to draw a parallel between finding the optimal balance in maximising universality *v.* maximising information power and finding the optimal balance in what are called in connectionist AI *underfitting & overfitting*. We say that we have an *overfitting* when that AI model fits too precisely to the data based on which it was trained, but it fails to generalise to new cases (Russell et al. 2021, p.673). E.g., MODEL II fails to generalise to cases of violating ARTICLE 9 *contra* MODEL I since MODEL II was designed to fit the very particular case of ARTICLE 10 violations. In juxtaposition to overfitting, we say that we have an *underfitting* of an AI model when that model fail to capture specific patterns in the training set outputting both the cases that it should include and cases that it should not (*ibid.*). E.g., MODEL I includes both the violation of ARTICLE 10, but also the violation of ARTICLE 2 (RIGHT TO LIFE) which is not violated when a state bans the publication and dissemination of books. In other words, overfitting is when the engineers overmaximise *informational power*, while underfitting is when they overmaximise *universality*.

- *Unification* (or *gluing*): The requirement of *unification* is about "*asking whether or not [an] explication permits practitioners working in different areas of philosophy and science to conduct dialogue about [the explicatum] without equivocation*" (Taylor 2015, p.664 and §7.3). It is essentially the practice of *gluing* laid out in §3.1, ¶¶3-4 & Figure 2.

  Taylor 2015 construes UNIFICATION as an adequacy criterion separate from FRUITFULNESS and the rest three. However, if FRUITFULNESS is about the fruitful impact the *explicatum* has in other disciplines, then the ability of the *explicatum* to be used from systems of concept from other disciplines is by default a FRUITFULNESS requirement. Furthermore, Taylor rightfully argues that unification is inversely correlated with the informational power of the *explicatum*: "*maybe the only way [an] impressive unification can be achieved is by making the explication almost content-free.*" (p.667,§8.3), i.e., making it *information-free* or at least reducing its informational content. The lesser the informational content the lesser the chances of contradictory systems of concepts and hence the higher the compatibility of those systems.[34] Finally, we should highlight the inherent conflict between the *parsimony* requirement and *unification*: unifying systems of concepts raises the overall ontology as well as the metaphysical, theoretical, methodological,

[34]By *compatibility with a system of concepts*, I mean minimally to *not contradict* the rules of use of those concepts and maximally to *reinforce* them. E.g., we will see in §IV.1 that the ECtHR rejects a specific subsumptive test for CAUSAL JUSTIFICATION, the so-called but-for test. Subsequently, the ECtHR's system of concepts can not be unified with systems of concepts where CAUSAL JUSTIFICATION includes the but-for test in its rules of use.

and epistemic commitments one makes since every system of concepts carries its own such commitments.

- *Maximising understandability & explanatory power:* Although in everyday discourse they are usually construed as synonyms, in philosophy of explanation, UNDERSTANDING and EXPLANATION are notoriously considered to be two distinct concepts (Grimm 2021, §4.1). For instance, a proof of a theorem is an *explanation* of why the theorem is true, but that does not entail that the reader of this proof will *understand* this explanation (*cf.* Mancosu, Poggiolesi, and Pincock 2023, §2). In other words, explanation can contribute to understanding, but understanding is *more* than a mere explanation (*see e.g.* Grimm 2021, §4.1; Siscoe 2022, §§1,3-4). That does not mean though that understanding is *per se* a special case of explanation since there are also those that argue that we can achieve understanding *without* an explanation (Hannon and Nguyen 2022, p.8).[35] A core difference between the two is that understanding seems to *depend* on the subjects that we want to understand, while explanations seem to be formed based on certain subject-*independent* standards (de Regt and Dieks 2005, §§1-2; *cf.* Hempel 1965; Trout 2002, 2005, 2007). UNDERSTANDING's subject-dependence is another element of FRUITFULNESS's user- and practice-dependence (§3.2.4, ¶1).

  understanding *v.* explanation

  Based on this difference, I construe *understandability* in a twofold way: (α) *expert*-oriented understandability: it is about making the concepts in the target systems understandable for the experts that explicate them in the target system (e.g., making PROBABILISTIC CAUSAL DIRECTED GRAPH understandable to legal experts or LEGAL CAUSAL JUSTIFICATION understandable to the AI engineers and logicians); (β) *user*-oriented understandability: it is about making a concept understandable for its users in its intended application (e.g., making the formal explication of CAUSAL JUSTIFICATION understandable to the ECtHR's judges (*cf.* Grimm 2021, §4.1)). This can be achieved by designing the appropriate *interface* between the user and the *explicatum* (*cf.* §I.3.2.1.1, ¶5). For a paradigmatic example of designing interfaces for AI models based on evaluations by domain experts *see* the ML-related paper Virgolin et al. 2021.

  Finally, I construe as *explanatory power* the ability of the *explicatum* to explain why it is used the way it is used. E.g., why the ECtHR can use CAUSATION to attribute legal responsibility for violations of human rights? Considering this, the requirements of *intentional similarity & coherence* can contribute positively to an *explicatum*'s explanatory power since they explicate under which conditions a concept is used, and many times they do so while also explaining *why* those are the appropriate conditions (*see* the demarcation between *genuine & non*-genuine rule of use in §3.3).

At this point, I end my adjustments of Brun's 2016 recipe of explication. More advanced research is required for further and more precise adjustments. Most of those adjustments will be fleshed out in the *actual* practice of explicating. It is neither realistic nor even recommended to decide *ex ante* every aspect of the practice of explication (*cf.* §3.2, ¶2). What needs to be decided is an initial foundation. And that foundation includes the decision of *how to begin* the explication. This is what I do in the next and last subsection of this CHAPTER. Afterwards, in CHAPTER IV, I will use this proposed methodology to begin the explication of CAUSAL JUSTIFICATION in the ECtHR's legal tradition of human rights law.

III.3.3 How to begin an explication

Before arguing how explication should start, it needs to be highlighted that in principle, the different steps of explication do *not* have to be followed *linearly*. They should be adjusted to the particularities of each case (Brun 2016, pp.1237-1238):

> "*...in practice, the process is non-linear and not rigidly structured for several reasons (cf. Stegmüller 1973, 25–36). Attempts at introducing an explicatum may prompt us to revise what we did in one of the 'previous' steps. It may turn out that the explicandum needs further disambiguation or we discover that we need more than one explicatum for different purposes. Or we find that the specified conditions of theoretical usefulness cannot be jointly satisfied and this may motivate us to stick to a certain explicatum and adapt the conditions of adequacy.[footnote] In some cases, the necessary clarification of the explicandum may call for identifying a subtle ambiguity which is most effectively identified indirectly by tentatively introducing explicata and comparing them with the help of the resources of the target system of concepts. Another reason is that explications give rise to feedback effects. A successful explication of a ordinary language concept can have the effect that the meaning of the explicandum-term changes or that the explicatum gets adopted into everyday language (as in the case of [fish]; see LaPorte 2003, ch. 3.IV).* ***A basically sequential structure of explicating has no room for these phenomena****, but an adequate account of conceptual reengineering needs to deal with them as well.*"
>
> Brun 2016, p.1237, emphasis added

[35] *See also* Lipton's 2009 "*Understanding without explanation*" *contra* Strevens's 2013 .

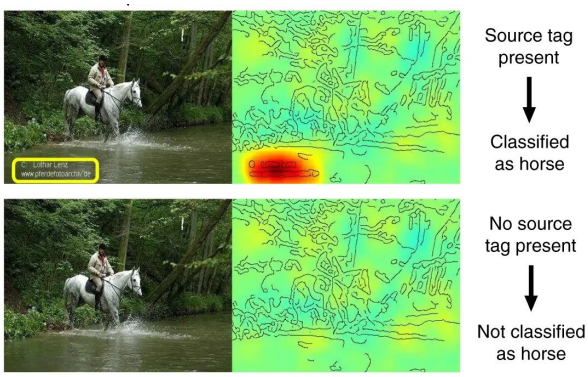
Having said that, I would still try to provide a schema of the basic steps that the explication engineers should start with and that they will have to repeat multiple times during the explication.

The goal of the conceptual engineers is to explicate CAUSAL JUSTIFICATION (the principal *explicandum*) as it is used in the ECtHR's system of concepts (the *source system* of concepts). I.e., they need to *express* in an *improved* way the rules of use of CAUSAL JUSTIFICATION in the target system of concept. Subsequently, the *first step* is to use the target system of concepts to express a tentative subset of the principal *explicandum*'s rules of use and *then* improve them. Since this is a CROSSDI practice, the principal *explicatum* will have to be expressed using the system of concepts of the *dominating* discipline. Afterwards, those rules of use will have to be *translated* to a (semi-)formal form so as to be incorporated in ALGOAI models. Consequently, the *second step* is to formalise the already rules of use introduced in the first step.

In both steps, we need to *introduce* concepts to the target system of concepts. By "*introducing a concept*", I mean identifying a subset of the concept's rules of use since those rules are what constitutes the concept (§2.2.1). So far, we have seen *three* ways of introducing concepts in the target system of concepts: (α) introducing what I will call *source concepts*. I.e., concepts from the source system of concepts like the principal *explicatum*. As already argued multiple times, since this is a CROSSDI practice, source concepts should be expressed *via* the dominating discipline's system of concepts. I.e., concepts like CAUSATION, JUSTIFICATION, RESPONSIBILITY and the like that are used by the ECtHR should be expressed by using the system of concepts of legal science. In other words, it is a case of *meta*-analysing the source system of concepts using the system of concepts of the dominating discipline.; (β) introducing what I will call *fruitful concepts*. I.e., concepts from the dominated disciplines that interact fruitfully with the source concepts so as to realise the TRANSDI ends. E.g., introducing the concept DIRECTED GRAPH so as to model CAUSAL JUSTIFICATION in order to realise the TRANSDI end of LEGITIMACY. Note that the *gluing concepts* of the fox-disciplines are a subset of the fruitful concepts.; (γ) introduce the TRANSDI ends by using the dominating discipline's system of concepts (*fn.* 18). As argued in §I.1.2, ¶¶4-5, the introduction of TRANSDI ends like RULE OF LAW and HUMAN RIGHTS should be done *via* the checklists of their operational definitions.

introducing concepts

• STEP I: introducing source concepts



The *heatmaps* (*cf.* §II.4.2.2) show which pixels contribute to the classification of an image as *horse*: the more red the pixel the more its contribution. The lack of a substantial amount of red pixels leads to the non-classification of an image as horse (Lapuschkin et al. 2019, Figure 2).

We saw that in the first step, we need to introduce source concepts by identifying their rules of use in the source system of concepts. To do so, it is imperative to identify criteria that will allow us to demarcate between *genuine* and *non-genuine* rules of use of the *explicanda*. A first demarcation is that between *coincidental* and *non*-coincidental rules of use. As *coincidental*l rules of use, I construe the non-genuine rules of use whose application leads to the same extension as the extension of the *explicandum* by *coincidence* and not by any relation to the *meaning* of the *explicatum*. It just happened that the two extensions *coincide*. Let's see a classical example of a coincidental rule of use in contemporary state-of-the-art connectionist AI. Lapuschkin et al. trained AI to successfully judge whether an image contains a horse or not (Lapuschkin et al. 2019, §E). Hence, the AI could successfully identify the extension of HORSE. However, it turned out that the said AI was performing this successful classification not because it had any sort of understanding of what a visual representation of horses is, but because the images with horses with which it was trained contained the same *watermark* (*ibid.*; *see also* Figure **??**). Ergo, in reality, the AI was identifying images with *watermark x* and not images with horses. It just happened that the extension of HORSE was *coinciding* with the extension of WATERMARK X. This is an example of what is called in the literature the *Clever Hans* phenomenon (hence another name for coincidental rules can be *Clever Hans rules*). Clever Hans was an Arabian stallion thatlived in 1890 Germany that was allegedly answering mathematical questions by tapping his front foot. In reality, the horse was not performing any mathematical calculations, but it was responding to physical cues performed by the auditor (Baskerville 2010).[36] In other words,

genuine *v.* *non*-genuine rules

Clever hans rules

[36]For the original findings debunking the "*math horse*" claims *see* Pfungst 1911.

despite the horse outputting a correct extension, it did not have any understanding of mathematical concepts.

Note that Clever Hans rules score really well with certain adequacy criteria. Assume for instance that we already know that the *explanandum*'s extension is the collection of the objects o_1, o_2, ..., o_n. E.g., assume that we already know which are the ECtHR's causal justifications. Then, the rule "*An object belongs to* CAUSAL JUSTIFICATION*'s extension if and only if it is either* o_1 *or* o_2 *or ... or* o_n." scores really high with EXTENSIONAL SIMILARITY, SIMPLICITY, and EXACTNESS. However, it scores negatively with INTENTIONAL SIMILARITY since we have no information about sufficient or necessary conditions for when a particular o is subsumed by the extension of CAUSAL JUSTIFICATION. It is the inclusion of such necessary and sufficient conditions that would demarcate a coincidental from a non-coincidental rule. For instance, a genuine binary classification of images to images with horses and images without horses presupposes that there must be certain visual characteristics that characterise horses either as necessary conditions for a horse to be a horse or as sufficient conditions for a particular to be horse. It is *only after* we have knowledge about those conditions that we have *genuine* understanding of the concept HORSE. Ergo, requirements of extensional similarity *should be grounded* on requirements of intentional similarity: similar necessary and sufficient conditions of a concept's application lead to similar applications of that concept, similar applications of a concept lead to similar extensions of that concept, and similar extensions of a concept lead to EXTENSIONAL SIMILARITY. If we try to reason the other way around (from EXTENSIONAL SIMILARITY to INTENTIONAL SIMILARITY), we are in danger of being accused of generating non-genuine coincidental rules.

grounding extension on intension

The question now becomes *how* can we explicate a concept's intention while staying loyal to its extension. A strategy to achieve this could be the coherence method of *narrow reflective equilibrium* (*RE*) that we saw in §3.2.1: we choose certain *paradigmatic examples* of CAUSAL JUSTIFICATION's extensions and we try to identify rules that explain more *adequately* those paradigmatic examples. *Adequacy* in this case can be construed in terms of explication's adequacy criteria. This way, we combine both conceptual engineering methods of explication & RE, as Brun 2020 suggests so as to balance out their deficiencies (*cf.* §3.2.1). Note that paradigmatic examples of the application of a concept include both *positive* and *negative* (Brun 2016, p.1227). E.g., we can identify rules of use of CAUSAL JUSTIFICATION by identifying paradigmatic examples of the application of the concept NON-CAUSAL JUSTIFICATION. If, for instance, a rule of use leads to the subsumption of a non-causal justification by CAUSAL JUSTIFICATION, then we should disregard that rule. Let's see now a *RE inference schema* proposed by Schroeter, Schroeter, and Toh 2020 (§VIII; *cf.* Dworkin 1986, pp.65-68; 2011, p.131; Stavropoulos 1996), as well as how it is bound with explication:

reflective equilibrium & explication

- RE STEP I: identify the *practice* associated with the source system. For Schroeter, Schroeter, and Toh, such practices can be legal traditions (2020, §IX). In our case, it is the legal tradition of the ECtHR. RE STEP I is the equivalent of identifying the source system of concepts in explication.
- RE STEP II: identify the practice's *core purposes*. This step is essentially the identification of the paradigmatic cases of the applications of the *explicanda*.
- RE STEP III: identify the best *realisers* (or *satisfiers* if you prefer) of those purposes. One can use the explication requirements to identify which are the best realisers of those purposes.

Note that Brun 2016 directly refers to STEPS I & II in his construal of explication arguing that SIMILARITY "*must be characterized by a* ***specification of the contexts*** *in which and the* ***purposes*** *for which the explicatum can replace the explicandum; that is, perform the* ***explicandum's function.***" (pp.1221-1222, emphasis added). For more similarities between RE & explication like their non-linear structure *see* Brun 2020, §§5.3,6.

Let's see now the particularities of explication's STEP II: *formalising source concepts introduced in* STEP I.

• STEP II: formalising source concepts

Clever Hans in flesh & blood!

We saw that in the step of formalisation we need to introduce fruitful concepts from the AI disciplines and glue them with the source concepts of legal science. This is where the foxes come to the rescue determining which are the appropriate formal tools to formalise the *source concepts* that the legal scientists explicated in natural language. Essentially, the process of formalisation is to *translate* in formal languages the rules of use of the source concepts. To evaluate the adequacy of such translations, based on the SIMILARITY requirement of coherence and the method of RE, logicians & formal philosophers will have to test whether those formalisation can identify paradigmatic applications of the *explicanda*. Since real life case-law is a quite complex phenomenon, it is advisable to first use simple examples for the tests. The literature of formal philosophy of explanation is full of such examples philosophers use to make their arguments. We will see many such examples in the next CHAPTER. Finally, the choice of *formal language* as well as of the *formal forms* (*see* the SIMILARITY requirement of identifying a concept's formal form) will be guided by the TRANSDI end of engineering a legal ALGOAI model. I.e., we should choose a formal form depending on the particularities of the AI model that the AI engineers will built. E.g., will they use logic-based models? And if yes what kind of logic will they use? What about Bayesian networks or probabilistic logical programming? And so forth.

In the next CHAPTER, I show how the introduction of formal rules of use in a formal form that is quite common in AI literature can be performed based on explication's STEPS I & II. I do so for the concept of CAUSAL JUSTIFICATION as used in the ECtHR's system of concepts.

References

Baker, Alan. 2022. "Simplicity." In *The Stanford Encyclopedia of Philosophy*, Summer 2022, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Barendt, Eric. 2019. "What is the harm of hate speech?" *Ethical theory and moral practice* 22 (3): 539–553. https://doi.org/10.1007/s10677-019-10002-0.

Baskerville, Jerry Ray. 2010. "Short report: What can educators learn from Clever Hans the math horse?" *Emergency Medicine Australasia* 22 (4): 330–331. https://doi.org/10.1111/j.1742-6723.2010.01308.x.

Beckers, Sander. 2021. "The counterfactual NESS definition of causation." In *Proceedings of the 35th AAAI conference on Artificial Intelligence*, 35:6210–6217. 7. AAAI Press. https://doi.org/10.1609/aaai.v35i7.16772.

Benacerraf, Paul. 1965. "What numbers could not be." *The Philosophical Review*, no. 1, 47–73.

Brun, Georg. 2016. "Explication as a method of conceptual re-engineering." *Erkenntnis* 81 (6): 1211–1241. https://doi.org/10.1007/s10670-015-9791-5.

———. 2020. "Conceptual re-engineering: from explication to reflective equilibrium." *Synthese* 197 (3): 925–954. https://doi.org/10.1007/s10670-015-9791-5.

Byron, Christine. 2016. "The European Court of Human Rights: a living instrument as applied to homosexuality." *The Judges' Journal* 55 (3): 36+. https://link-gale-com.proxy.uba.uva.nl/apps/doc/A467915272/AONE?u=amst&sid=bookmark-AONE&xid=6502fed9.

Cabalar, Pedro, Jorge Fandinno, and Michael Fink. 2014. "Causal graph justifications of logic programs." *Theory and Practice of Logic Programming* 14 (4-5): 603–618. https://doi.org/10.1017/S1471068414000234.

Cabalar, Pedro, Jorge Fandinno, and Brais Muñiz. 2020. "A system for explainable Answer Set Programming." *Electronic Proceedings in Theoretical Computer Science* 325:124–136. https://doi.org/10.4204/eptcs.325.19.

Carnap, Rudolf. 1962. *Logical foundations of probability*. 2nd ed. University of Chicago Press.

Carnap, Rudolf. 1963. "Replies and systematic exposition." In *The philosophy of Rudolf Carnap*, edited by Paul A. Schilpp, 11:858–1013. The library of living philosophers. Open Court.

Chalmers, David J. 2020. "What is conceptual engineering and what should it be?" *Inquiry: an interdisciplinary journal of philosophy*, 1–18. https://doi.org/10.1080/0020174X.2020.1817141.

Cook, Roy T. 2009. *A dictionary of philosophical logic*. Edinburgh University Press.

Daniels, Norman. 2020. "Reflective Equilibrium." In *The Stanford Encyclopedia of Philosophy*, Summer 2020, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Danks, David. 2014. "Learning." Chap. 4 in *The Cambridge handbook of artificial intelligence: Responsible artificial intelligence*, edited by Keith Frankish and William M. Ramsey, Part III: Dimensions. Cambridge University Press.

de Regt, Henk W., and Dennis Dieks. 2005. "A contextual approach to scientific understanding." *Synthese* 144 (1): 137–170. https://doi.org/10.1007/s11229-005-5000-4.

Dean, Walter. 2021. "Computational complexity theory." In *The Stanford Encyclopedia of Philosophy*, Fall 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Duchessi, Peter, and Robert M. O'Keefe. 1995. "Understanding expert systems success and failure." *Expert Systems with Applications* 9 (2): 123–133. https://doi.org/10.1016/0957-4174(94)00056-2.

Dutilh Novaes, Catarinar. 2020. "Carnapian explication and ameliorative analysis: a systematic comparison." *Synthese* 197 (3): 1011–1034. https://doi.org/10.1007/s11229-018-1732-9.

Dworkin, Ronald. 1986. *Law's empire*. Harvard University Press.

———. 2011. *Justice for hedgehogs*. Belknap Press.

ECtHR Registry. 2021. *Guide on Article 10 of the European Convention on Human Rights: Freedom of expression*. Updated. April. https://www.echr.coe.int/documents/guide_art_10_eng.pdf.

ECtHR's Press Unit (Unité de la Presse). 2023. *Factsheet on sexual orientation*. January. Accessed April 4, 2023. https://www.echr.coe.int/Pages/home.aspx?p=press/factsheets&c.

Fine, Kit. 1975. "Vagueness, truth and logic." *Synthese* 30 (3/4): 265–300.

Frost, Rebecca L. A., and Padraic Monaghan. 2020. "Insights from studying statistical learning." In *Current perspectives on child language acquisition: how children use their environment to learn*, edited by Ben Ambridge Caroline F. Rowland Anna L. Theakston and Katherine E. Twomey, 27:65–89. Language acquisition research. John Benjamins Publishing Company. https://doi.org/10.1075/tilar.27.03fro.

Goodman, Nelson. 1955. *Fact, Fiction, and Forecast*. Harvard University Press.

———. 1958. "The test of simplicity." *Science* 128 (3331): 1064–1069.

Gordon, Thomas F. 1988. "The importance of nonmonotonicity for legal reasoning." In *Expert systems in law: Impacts on legal theory and computer law*, edited by Herbert Fiedler, Fritjof Haft, and Roland Traunmüller, 4:111–126. Neue methoden im Recht. Attempto Verlag Tübingen GmbH.

Governatori, Guido, Antonino Rotolo, and Giovanni Sartor. 2021. "Logic and the law: philosophical foundations, deontics, and defeasible reasoning." Chap. 9 in *Handbook of deontic logic and normative systems*, edited by Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, vol. 2. College Publications.

Griffin, James. 2008. *On human rights*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199238781.001.0001.

Grimm, Stephen. 2021. "Understanding." In *The Stanford Encyclopedia of Philosophy*, Summer 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Halpern, Joseph Y. 2016. *Actual causality*. The MIT Press.

Hannon, Michael, and James Nguyen. 2022. "Understanding philosophy." *Inquiry: An Interdisciplinary Journal of Philosophy*, 1–37. https://doi.org/10.1080/0020174X.2022.2146186.

Haslanger, Sally. 2012. *Resisting reality*. Oxford University Press.

Heinze-Deml, Christina, Marloes H. Maathuis, and Nicolai Meinshausen. 2018. "Causal structure learning." *Annual Review of Statistics and Its Application* 5 (1): 371–391. https://doi.org/10.1146/annurev-statistics-031017-100630.

Hempel, Carl G. 1965. *Aspects of scientific explanation and other essays in the philosophy of science.* Free Press.

———. (1988) 2000. "On the cognitive status and the rationale of scientific methodology." Chap. 11 in *Selected philosophical essays,* edited by Richard Jeffrey, 199–228. Cambridge University Press. https://doi.org/10.1017/CBO9780511815157.017.

Isaac, Manuel Gustavo. 2020. "How to conceptually engineer conceptual engineering?" *Inquiry: An Interdisciplinary Journal of Philosophy,* 1–24. https://doi.org/10.1080/0020174X.2020.1719881.

Jung Lee, Hyun. 2022. *Discrimination based on sexual orientation: Jurisprudence of the European Court of Human Rights and the Constitutional Court of Korea.* Interdisciplinary Studies in Human Rights. Springer Cham. https://doi.org/10.1007/978-3-030-95423-9.

Kalanov, Temur Z. 2013. "The critical analysis of the Pythagorean theorem and of the problem of irrational numbers." *Bulletin of Pure & Applied Sciences-Mathematics* 32E (1): 1–12.

LaPorte, Joseph. 2003. *Natural kinds and conceptual change.* Cambridge studies in philosophy and biology. Cambridge University Press. https://doi.org/10.1017/CBO9780511527319.

Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. "Unmasking Clever Hans predictors and assessing what machines really learn." *Nature Communications* 10 (1): 1096. https://doi.org/10.1038/s41467-019-08987-4.

Leitgeb, Hannes. 2013. "Scientific philosophy, mathematical philosophy, and all that." *Metaphilosophy* 44 (3): 267–275. https://doi.org/https://doi.org/10.1111/meta.12029.

Leitgeb, Hannes, and André Carus. 2022. "Rudolf Carnap." In *The Stanford Encyclopedia of Philosophy,* Fall 2022, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Letwin, Jeremy. 2021. "Why completeness and coherence matter for the European Court of Human Rights." *European Convention on Human Rights Law Review* 2 (1): 119–154. https://doi.org/10.1163/26663236-bja10002.

Lipton, Peter. 2009. "Understanding without explanation." In *Scientific understanding: Philosophical perspectives.* Edited by Henk W. de Regt, Sabina Leonelli, and Kai Eigner, Part I: Understanding, explanation, and intelligibility, 42–63. University of Pittsburgh Press.

MacCormick, Neil. 1992. "Legal deduction, legal predicates and expert systems." *International Journal for the Semiotics of Law* 5 (2): 181–202. https://doi.org/10.1007/BF01101868.

Machery, Edouard. 2017. *Philosophy withing its proper bounds.* Oxford University Press.

Mancosu, Paolo, Francesca Poggiolesi, and Christopher Pincock. 2023. "Mathematical Explanation." In *The Stanford Encyclopedia of Philosophy,* Fall 2023, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Margolis, Eric, and Stephen Laurence. 2022. "Concepts." In *The Stanford Encyclopedia of Philosophy,* Fall 2022, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Marmor, Andrei. 2018. "Varieties of vagueness in the law." In *Handbook of legal reasoning and argumentation,* edited by Giorgio Bongiovanni, Gerald Postema, Antonino Rotolo, Giovanni Sartor, Chiara Valentini, and Douglas Walton, Part III: Special kinds of legal reasoning, 561–580. Springer.

May, Tim. 2011. *Social research: Issues, methods and process.* 4th ed. Open University Press.

Mchangama, Jacob, and Natalie Alkiviadou. 2021. "Hate speech and the European Court of Human Rights: Whatever happened to the right to offend, shock or disturb?" *Human rights law review* 21 (4): 1008–1042. https://doi.org/10.1093/hrlr/ngab015.

Moore, Michael S. 2019. "Causation in the law." In *The Stanford Encyclopedia of Philosophy,* Winter 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

O'Shaughnessy, Matthew, Gregory Canal, Marissa Connor, Mark Davenport, and Christopher Rozell. 2020. "Generative causal explanations of black-box classifiers." In *Proceedings of the 34th International Conference*

on Neural Information Processing Systems, 5453–5467. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc.

Pfungst, Oskar. 1911. *Clever Hans (The Horse of Mr. Von Osten): A contribution to experimental animal and human psychology*. 2010 open-access ed. by Project Gutenberg. Translated by Carl Leo Rahn. Henry Holt and Company. https://www.gutenberg.org/ebooks/33936.

Poggi, Francesca. 2021. "Defeasibility, law, and argumentation: A critical view from an interpretative standpoint." *Argumentation* 35 (3): 409–434. https://doi.org/10.1007/s10503-020-09544-w.

Poscher, Ralf. 2012. "Ambiguity and vagueness in legal interpretation." Chap. 9 in *The Oxford Handbook of language and law*, Part II: The interpretation of legal texts, 128–145. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199572120.013.0010.

Prakken, Hendrik (Henry). 1993. "Logical tools for modelling legal argument." PhD diss., Vrije Universiteit.

Prinzing, Michael. 2018. "The revisionist's rubric: Conceptual engineering and the discontinuity objection." *Inquiry: An Interdisciplinary Journal of Philosophy* 61 (8): 854–880. https://doi.org/10.1080/0020174X.2017.1385522.

Rawls, John. (1971) 2005. *A theory of justice*. Reprint. Belknap Press.

———. 1999. *A theory of justice*. Revised ed. Belknap Press.

Rigoni, Adam W. 2014. "Legal rules, legal reasoning, and nonmonotonic logic." PhD diss., University of Michigan.

Russell, Stuart J., Peter Norvig, Ming-Wei Chang, Jacob Devlin, Anca Dragan, David Forsyth, Ian Goodfellow, et al. 2021. *Artificial Intelligence: A modern approach*. 4th ed. Gobal ed. Edited by Stuart Russell and Peter Norvig. Pearson series in artificial intelligence. Pearson.

Russo, Federica. 2022. *Techno-scientific practices: An informational approach*. Rowman & Littlefield Publishers.

Sartor, Giovanni. 2012. "Defeasibility in legal reasoning." Chap. 6 in *The logic of legal requirements: Essays on defeasibility*, edited by Beltrán Jordi Ferrer and Ratti Giovanni Battista, Part I: General features of defeasibility in law and logic, 108–136. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199661640.003.0007.

Schauer, Frederick. 1991. *Playing by the rules: A philosophical investigation of rule-based decision-making in law and life*. Clarendon Law Series. Oxford University Press.

Schroeter, Francois, Laura Schroeter, and Kevin Toh. 2020. "A new interpretivist metasemantics for fundamental legal disagreements." *Legal Theory* 26 (1): 62–99. https://doi.org/10.1017/S1352325220000063.

Schumann, Walter. 2008. *Minerals of the world*. 2nd ed. Sterling.

Siscoe, Robert Weston. 2022. "Grounding, understanding, and explanation." *Pacific Philosophical Quarterly* 103 (4): 791–815. https://doi.org/10.1111/papq.12391.

Smaldino, Paul E. 2017. "Models are stupid, and we need more of them." Chap. 14 in *Computational social psychology*, edited by Robin R. Vallacher, Stephen J. Read, and Andrzej Nowak, 311–331. Routledge. https://doi.org/10.4324/9781315173726.

Sorensen, Roy. 2022. "Vagueness." In *The Stanford Encyclopedia of Philosophy*, Winter 2022, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Spaak, Torben. 2009. "Explicating the concept of legal competence." In *Concepts in law*, edited by Jaap C. Hage and Dietmar von der Pfordten, 88:67–80. Law and Philosophy Library. Springer.

Speaks, Jeff. 2021. "Theories of meaning." In *The Stanford Encyclopedia of Philosophy*, Spring 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Stanford, P. Kyle. 2015. ""Atoms exist" is probably true, and other facts that should not comfort scientific realists." *The Journal of Philosophy* 112 (8): 397–416.

Stavropoulos, Nicos. 1996. *Objectivity in law*. Oxford University Press.

Stegmüller, Wolfgang. 1973. "Jenseits von Popper und Carnap." In *Jenseits von Popper und Carnap 'Stützungslogik, likelihood, bayesianismus statistische daten zufall und stichprobenauswahl testtheorie schätzungstheorie subjek-*

tivismus kontra objektivismus fiduzial-wahrscheinlichkeit, 15–60. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-52178-2_2.

Strevens, Michael. 2013. "No understanding without explanation." *Studies in History and Philosophy of Science Part A* 44 (3): 510–515. https://doi.org/10.1016/j.shpsa.2012.12.005.

Taylor, Elanor. 2015. "An explication of emergence." *Synthese* 172 (3): 653–669. https://doi.org/10.1007/s11098-014-0324-x.

Trout, J. D. 2002. "Scientific explanation and the sense of understanding." 69:212–233.

———. 2005. "Paying the price for a theory of explanation: de Regt's discussion of Trout." *Philosophy of Science* 72:198–208.

———. 2007. "The psychology of scientific explanation." *Philosophy Compass* 2 (3): 564–91.

Vedder, Richard G., Thomas P. van Dyke, and Victor R. Prybutok. 2002. "Death of an expert system: A case study of success and failure." *Journal of International Information Management* 11 (1).

Virgolin, Marco, Andrea De Lorenzo, Francesca Randone, Eric Medvet, and Mattias Wahde. 2021. "Model learning with personalized interpretability estimation (ML-PIE)." In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 1355–1364. GECCO '21. Lille, France. https://doi.org/10.1145/3449726.3463166.

Walton, Douglas. 2002. *Legal argumentation and evidence.* The Pennsylvania State University Press.

Webley, Lisa. 2010. "Qualitative approaches to empirical legal research." In *The Oxford handbook of empirical legal research*, edited by Peter Cane and Herbert M. Kritzer, 926–950. Oxford Handbooks. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199542475.001.0001.

CHAPTER IV

Going META
Explicating ECtHR's causal justifications

In this chapter, I apply the schema of explication laid out in CHAPTER III to the concept of CAUSAL JUSTIFICATION as used in the source system of the ECtHR's practice. As already argued in the introduction of CHAPTER III, the goal is to exhibit in a toy example how explication can be applied and *not* to provide a full-fledged account of an explicated CAUSAL JUSTIFICATION. The proposed semi-formal explication is Beckers's 2021 formal explication of: (α) Wright's 1985; 1988; 2011 concept NESS CAUSATION; (β) Hart's and Honoré's 1985 notorious BUT-FOR CAUSATION (or SINE QUA NON CAUSATION). More precisely, I show how according to the explication schema of CHAPTER III, Beckers's explications of NESS & BUT-FOR CAUSATION can be used to explicate ECtHR's CAUSAL JUSTIFICATION. Once more, I annotate concepts by using SMALL CAPS. I further use SMALL CAPS for explication's jargon (e.g., FRUITFULNESS, CORE PURPOSES, EXTENSIONAL SIMILARITY). Finally, note that multiple aspects of the theory fleshed out in the previous chapters will be present *implicitly* and *not* explicitly in this chapter as required by the EXACTNESS requirement of EXPLICITNESS *v.* IMPLICITNESS.

IV.1 STEP I: introducing SOURCE CONCEPTS

We saw in §III.3.3, that the first step of an explication should be to identify rules of use of the PRINCIPAL EXPLICANDUM using legal science's SYSTEM OF CONCEPTS. Inevitably, we will have to introduce further SOURCE CONCEPTS so as to articulate the PRINCIPAL EXPLICANDUM's rules of use.

The PRINCIPAL EXPLICANDUM in this case is the concept CAUSAL JUSTIFICATION. In the literature of legal science, one can find multiple terms referring to that concept: "*causal explanation*", "*causation*", "*legal causation*", "*causality*", etc. In other disciplines, those terms can refer to different concepts. E.g., we will see in §IV.2, ¶4 that in (formal) philosophy of explanation CAUSAL JUSTIFICATION differs from CAUSAL EXPLANATION. In the context of legal science though, in principal, they are used to satisfy the same CORE PURPOSE: *justifying* judicial judgements. Since they are *used* for the same CORE PURPOSE in the same PRACTICE, they have more or less the *same* rules of use. I.e., in this context, they can be construed as EXTENSIONALLY EQUIVALENT (or at least as EXTENSIONALLY SIMILAR). For reasons of UNIFICATION, PARSIMONY & UNDERSTANDABILITY (both *USER-* & *EXPERT*-ORIENTED UNDERSTANDABILITY), I will use only *one* term to notate the common rules of use of those concepts: "*causal justification*".

CORE PURPOSES & EXTENSIONAL EQUIVALENCE

A PARADIGMATIC example of justifying judgements *via* CAUSAL JUSTIFICATION is the use of CAUSAL JUSTIFICATION to attribute *responsibility* for the violation of the law (Hart and Honoré 1985, §III; So 2020, §2.2; Shafer 2002, §1.3; Moore 2009, §I; Wright 1988, §§III,VI; *cf.* Sartorio 2009; Halpern 2016, §6). For instance, in the cases where a court has to decide whether a defendant is *responsible* for the victim's death, if there is a causal relation between certain acts[1] of the defendant and the death of the victim, then there are high chances that the court will hold the defendant responsible. We do not care about CAUSAL JUSTIFICATION in any legal practice though, but in the particular practice of the ECtHR. The ECtHR *also* employs CAUSAL JUSTIFICATION so as to attribute responsibility, albeit in an incoherent and inconsistent way (Stoyanova 2018 *contra* Lavrysen 2018; for more literature on causation in the ECtHR legal tradition *see also* Turton 2020; Stoyanova 2020; Sulyok 2017; Nolan 2013). The existing literature on *formal* accounts of CAUSATION in the ECtHR's practice is almost non-existent *contra* to the literature on the Anglo-American legal tradition. Novel attempts of explicating causation in other legal traditions usually attempt to apply the Anlgo-American formalism to their tradition. For reasons

CORE PURPOSE: responsibility attribution

[1] I refer to both *positive* and *negative* acts like death by shooting (*see e.g.* Shafer 2002, §1.3.2) and death by negligence (*see e.g.* Green 2015; Turton 2020) respectively. For the definitions of positive and negative acts *remember* §I.1.1, ¶1; *cf.* §I.1.2, ¶5.

of UNIFICATION, my approach will follow the same strategy.

Not so surprisingly if we consider the importance of CAUSAL JUSTIFICATION in attributing responsibility, contemporary PARADIGMATIC (formal) explications of CAUSATION stem from explications of CAUSAL JUSTIFICATION, CAUSATION, CAUSALITY and the like in *law* (*cf.* §2, ¶1). The standard reference point is Hart's & Honoré's 1959 (1st ed.) and 1985 (2nd ed.) "*Causation in the law*" (*cf.* Wright 2011, §II; Summers 2018). Hart's & Honoré's explication of the use of CAUSATION in legal practice was that of the SINE QUA NON CAUSATION or as it has become known the BUT-FOR CAUSATION, "*[t]he most widely used test [by judicial authorities] – and most vehemently criticized...*" (Plakokefalos 2015, p.476). Despite BUT-FOR CAUSATION's insufficiency to model PARADIGMATIC cases of causation used for responsibility attribution like OVERDETERMINED CAUSATION that we will be explicated later on, BUT-FOR CAUSATION "*remains at the core of most legal inquiries*" (Summers 2018, p.795,footnote 7; *see also* Green 2015, §2). So, what is BUT-FOR CAUSATION and how is it used for responsibility attribution? The answer is given by Hart's & Honoré's *bifurcation of the causal question* (1985, p.110):

THE CAUSAL QUESTION: Was harm $\mathcal{E}$ the consequence of act $\mathcal{C}$?;

THE FACTUAL BRANCH: Would the harm $\mathcal{E}$ have occured if act $\mathcal{C}$ had not occurred?;

THE LEGAL BRANCH: Is there any principle which precludes the treatment of $\mathcal{E}$ as the consequence of $\mathcal{C}$ for *legal purposes*?

Based on this bifurcation, I will introduce in the next paragraph the concepts FACTUAL CAUSATION & LEGAL CAUSATION, and right after, based on those two concepts, I will introduce BUT-FOR CAUSATION and its use by judicial authorities to attribute responsibility.

Hart and Honoré's bifurcation of the causal question is essentially two SUBSUMPTION TESTS: whenever the FACTUAL BRANCH is answered negatively, then we have a case of what is called in the literature of jurisprudence as FACTUAL (or NATURAL) CAUSATION with $\mathcal{C}$ being the *factual* cause & $\mathcal{E}$ the *factual* effect. While whenever both branches are answered negatively, the we have a case of what is called in the literature of jurisprudence as LEGAL CAUSATION with $\mathcal{C}$ being the *legal* cause & $\mathcal{E}$ the *legal* effect (for a comparative analysis between FACTUAL CAUSATION & LEGAL CAUSATION in tort law[2] from different legal traditions have a look at Koziol 2015; *see also* Askeland 2015, §II.A,p.125; Stoyanova 2018, pp.940,942; UN ILC 2001, p.39; Koziol 2015, p.812). *Ergo*, *legal* causal relations are a subset of *natural* causal relations. I.e., LEGAL CAUSATION is a HYPONYM of FACTUAL RELATION and legal causal relations are subsumed by FACTUAL CAUSAL RELATION. As we saw in §III.3.2.1, establishing a relation of hyponym between the concepts of LEGAL & FACTUAL CAUSATION is a realisation of RELATIONAL SIMILARITY. Now the extra conditions used to choose which factual causal relations constitute *legal* causal relations are relevant to each legal tradition & area of law. For instance, according to the Hungarian Civil Code 2013, a necessary condition for determining whether a *factual causal relation* between a cause and its effect is also a *legal* causal relation is for the effect to be *forseeable* given its *cause* (Menyhárd 2015, p.297) showing once more the importance of *forseeability* even in illiberal democracies (*cf.* §I.2.8, ¶2). Identifying such *necessary* conditions for determining whether a causal relation is subsumed by LEGAL CAUSAL RELATION is a realisation of the INTENTIONAL SIMILARITY requirement. Let's introduce not BUT-FOR CAUSATION using our knowledge about FACTUAL & LEGAL CAUSATION.

RELATIONAL SIMILARITY

INTENTIONAL SIMILARITY

What was explicated in the previous paragraph as FACTUAL CAUSATION is what in contemporary causal inference literature is called BUT-FOR CAUSATION and what Hart and Honoré 1985 called SINE QUA NON CAUSATION. I.e., if we know that *but-for* $\mathcal{C}$, $\mathcal{E}$ would not have occured, then we can characterise $\mathcal{C}$ as the *but-for cause* of the *but-for effect* $\mathcal{E}$.[3] According to Hart and Honoré 1985, BUT-FOR CAUSATION is necessary for attributing legal responsibility. More precisely, an agent[4] should be held legally responsible for a harm $\mathcal{E}$ whenever their act $\mathcal{C}$ is a *legal cause* of $\mathcal{E}$. I.e., to determine whether the agent is responsible we need to first perform the *but-for test*, and then, if it is successful, we need to see whether the rest of LEGAL CAUSATION's subsumptive tests are satisfied (e.g., testing whether the effect was foreseeable given the cause as is the case in Hungarian tort law (*see* previous paragraph)). Hart's & Honoré's normative suggestion of how BUT-FOR CAUSATION should be used to attribute responsibility is a standard way of attributing responsibility in many *actual* judicial practices like the case of the Hungarian Civil Code 2013 cited above. After all, all cited authors from the literature of jurisprudence (Hart and Honoré; Wright; Moore; Stoyanova; Green, etc) explicate what happens in the *actual* legal practice or how

BUT-FOR CAUSATION

[2] *Tort law* "*is mainly concerned with providing compensation for personal injury and property damage caused by *negligence*" (Law 2022).

[3] Note though that as we will see in the next subsection (§1.1), not every legal scientist & judicial authority agrees with this explication of FACTUAL CAUSATION. The intuition behind the concept of FACTUAL CAUSATION that formal philosophers of causation try to explicate is that it refers to causal relations that exist in the world *independently* of any condition imposed by legal orders (Hamer 2014, pp.155-156).

[4] Remember the relation between *agents* and *legal responsibility* as well as why ALGOAI should not be held liable for misuse/abuse of power: §I.2.1, ¶5; §I.3.2.1.1, ¶7.

that practice should be reformed. In other words, they use *empirical* data to make their points as legal scientists do (*cf.* §II.4.1.2). For instance, the but-for test can be found already from 1962 at American Law Institute's Model Penal Code §2.03, ¶11 (Proposed Official Draft, 1962, emphasis added) which has been used extensively in the US criminal law judicial practice[5] (*cf.* Moore 2009, pp.87,167; for the use of the but-for test by the Anglo-American legal tradition *see also* Moore 2009, 2019; Green 2015, §1; Robertson 2009, pp.1008-1009,footnotes 11-15):

> **Causal relationship between conduct [CAUSE] and result [EFFECT]**
>
> A conduct is the cause of a result when:
>
> (a) it is an antecedent *but for* which the-result-in question would not have occurred; [THE FACTUAL BRANCH]
>
> (b) the relationship between the conduct and result satisfies any additional causal requirements imposed-by-the Code or by the law defining the offense [THE LEGAL BRANCH]

Despite its importance in responsibility attribution, it should be noted that the but-for causation is *not* a *necessary* condition for an agent to be held legally responsible for a harm. E.g., a parent can be found legally responsible for a harm caused not by them but by their child. Or a generation of agents may be found responsible for a harm caused by a previous generation of agents (e.g., Germany compensating Israel for WWII war crimes) (Hart and Honoré 1985, pp.63-64).

What about the legal practice of the ECtHR though? Is the but-for test used for the CORE PURPOSE of responsibility attribution?

IV.1.1 The ECtHR & the problems with the but-for subsumptive test

rejecting the but-for test

The ECtHR has *explicitly* rejected the but-for test as a test that attributes responsibility. More precisely, in the *E. and others v. UK* case, the applicants argued that the lack of investigation, communication, and cooperation of UK state-actors contributed to their harm (trauma from domestic abuse) violating ARTICLE 3 (PROHIBITION OF TORTURE). However, there was no sufficient evidence to support that indeed, *but-for* the UK state-actors lack of response, the harm would have been avoided. I.e., there was not enough evidence to support the but-for test (*cf.* Turton 2020). Despite this inability to establish a causal *nexus*, the Court judged that:

> "*... Article 3 however does* ***not*** *[require] to be shown that* ***"but for"*** *the failing or omission of the public authority ill-treatment would not have happened. A failure to take reasonably available measures which could have had a real prospect of altering the outcome or mitigating the harm is sufficient to engage the responsibility of the State.*"
>
> *E. and others v. UK (2002)*, ¶99, emphasis added.

Subsequently, the Court found the UK *responsible* for the applicants' harm since UK authorities should have tried to protect the applicants' human rights as long as there was "*a real prospect of altering the outcome or mitigating the harm* " even if there was no evidence to suggest that doing so would have lead to a different outcome.

The foregoing example is not an exceptional case in the ECtHR's case-law. There are multiple cases in which the ECtHR attributes responsibility to HCPs based on what Stoyanova 2018 called "*domestic legality*" (§6.A): if there are domestic laws in place that protect human rights and the HCP did not follow them, even if there is no evidence to suggest that following them would have decisively prevented or mitigated the harm, the HCP is still *responsible* for that harm (*see e.g. Elena Cojocaru v. Romania (2003), I. v. Finland (2008), Lopes de Sousa Fernandes v. Portugal (2017)*). Having said that, BUT-FOR CAUSATION can still appear in those justifications like in the *E. and others v. UK* quote above. In that quote, the Court justifies to the involved parties *why* BUT-FOR CAUSATION is not good enough to attribute responsibility. A justification after all has to include the arguments that judicial authorities *rejected* justifying *why* they were rejected. Moreover, due to the *principle of equality*,[6] the same justification will be used in future cases to reject similar arguments. And indeed, the Court has used

[5] The Model Penal Code (MPC), as a *code*, it is a systematised clarification of existing laws (Black 1968, p.323). Since it is a *penal* code, it is about *criminal* law, US criminal law in this case. And "*model*" means that it is a *proposal* of how criminal law should be applied. In the case of MPC, the majority of the US states have modeled their penal codes based on it (*West's Encyclopedia of American Law* 2004). It has been extensively used by US courts in their practice and it has been a standard textbook for teaching criminal law to undergraduates (Dubber 2015, §1). Ergo, this is another example of the wide use of BUT-FOR CAUSATION in the attribution of legal responsibility by *actual* judicial authorities. It is also worth mentioning that the intellectual origins of the MPC are attempts to clarify the law by systematising its application using *philosophical* tools. I.e., it is another type of *conceptual re-engineering* which in this case originated from utilitarian philosopher and British jurist Jeremy Bentham (1748 -1832) (Kadish 1978, p.1099; *see also* Crimmins 2021).

[6] "*... the law [should] treat similar situations similarly and dissimilar situations dissimilarly*" (§I.2.4, ¶2 ; *cf.* CDL-AD(2016)007, p.18; §III.2.2.2).

it many times raising controversy in HCPs (e.g., in the UK) that find the rejection of the but-for test a very *loose* condition for responsibility attribution (Turton 2020). Considering the above, ALGOAI justifications of future cases with similar characteristics to the *E. and Others v. United Kingdom* case should also justify why the but-for test is rejected. This a case of a NEGATIVE EXAMPLE of the concept BUT-FOR CAUSATION, and as argued in §3.3, ¶5, NEGATIVE EXAMPLES should still be taken into consideration during the process of explication.

NEGATIVE EXAMPLE

The fact that the ECtHR does not consider a successful but-for test as a *necessary* condition for responsibility attribution does not preclude the possibility that under certain circumstances BUT-FOR CAUSATION is indeed used for responsibility attribution. Thus, we need to inquire whether this is the case. To do so, we should disambiguate between FACTUAL CAUSATION and BUT-FOR CAUSATION. As argued in *fn.* 3, while BUT-FOR CAUSATION is an *explication* of FACTUAL CAUSATION proposed by legal scientists like Hart and Honoré and used by actual judicial authorities like Hungarian civil courts (§1, ¶6), that does not mean that every legal scientist and every judicial authority agrees with this explication. The ECtHR, seems to still use the bifurcation between FACTUAL & LEGAL CAUSATION to attribute responsibility, albeit not always with the but-for explication of FACTUAL CAUSATION.[7] Let's see first an example of the but-for explication of FACTUAL CAUSATION used by the Court for responsibility attribution, and then, we will see examples where it is ill-fit to explicate. The but-for example I will use is that of the *Mastromatteo v. Italy (2002)* case since *prima facie* it seems to reject the but-for test, albeit this is not true. The facts of the case are that Italian authorities granted prisoners M.R. & G.M. permission to exit prison (prison leave & semi-custodial treatment respectively) and the prisoners ended up murdering the applicant's son. As one can see from the quote below, the Court judged that in this case, the but-for test for the FACTUAL CAUSATION holds: *but-for* M.R. & G.M. not being in jail, A. Mastromatteo would not have been murdered. However, the Court found that legal principles that would make this *factual* causal relation a *legal* causal relation (e.g., the expectation of M.R & G.M. murdering A. Mastromatteo being *reasonable* and his murder being considered a *real and immediate risk* (*cf.* Stoyanova 2018, §§7.A-7.B)) do not hold, and ergo, the Italian state should not be found responsible. Had these conditions be satisfied, e.g. had the expectation of the A. Mastromatteo being *reasonable* and had his murder being a *real and immediate risk*, then we would have a case of LEGAL CAUSATION and of a legitimate responsibility attribution.[8] One can construe this as a *weakening* of the but-for test and hence why Turton 2020 calls it a *loose* but-for test.

BUT-FOR CAUSATION ≠ NESS CAUSATION

> *"... it is clear that if M.R. and G.M. had been in prison on 8 November 1989, A. Mastromatteo* ***would not have been murdered*** *by them. However,* ***a mere condition sine qua non*** *does* ***not*** *suffice to engage the responsibility of the State under the Convention; it must be shown that the death of A. Mastromatteo resulted from a failure on the part of the national authorities to "do all that could* ***reasonably*** *be expected of them to avoid a* ***real and immediate risk*** *to life of which they had or ought to have had knowledge" [Osman v. UK, ¶116 (1998)],..."*
>
> *Mastromatteo v. Italy*, ¶74, emphasis added.

Let's see now why the but-for test fails to capture every case of FACTUAL CAUSATION in the ECtHR's practice as well as which subsumptive test could make up for those deficiencies. A significant drawback of the but-for test is that it fails to identify causal relations when we have an OVERDETERMINATION of causes (aka (ACTUAL) DUPLICATIVE CAUSATION (Wright 1985, §II.E.2; Green 2015, §4)). CAUSAL OVERDETERMINATION is when there is more than one cause that can sufficiently cause by the effect just by itself. The classical example in the literature of formal philosophy of causation is that of Bob and Patrick each throwing a stone that hit a bottle at the same time breaking it. Both of them are causes of the bottle breaking, but both fail the but-for test. *But for* Bob not throwing the rock, the bottle would *still* have broken due to Patrick throwing another rock. And similarly, *but for* Patrick not throwing the rock, the bottle would still have broken due to Bob throwing another rock. Another example of overdetermination is the example of voting. If 8 people out of the 10 vote to go to a beach, while all 8 of them contributed to going to the beach, none of them is a but-for cause (*cf.* Halpern 2016, Example 2.3.2). Note that causal overdetermination is quite common in law (*see e.g.* Plakokefalos 2015 for the problem of attributing legal responsibility to *states* in cases of CAUSAL OVERDETERMINATION). It is mainly due to causal overdetermination in law that Richard W. Wright critisised Hart's and Honoré's 1985 explication of FACTUAL CAUSATION as BUT-FOR CAUSATION ending up proposing its replacement by what they called NESS CAUSATION: *necessary element of a sufficient condition* (Wright 1985, 1988, 2011).

BUT-FOR CAUSATION *v.* overdetermination

How does NESS CAUSATION solve the problem of CAUSAL OVERDETERMINATION? The subsumptive test for NESS CAUSATION is that $\mathcal{C}$ is the NESS cause of $\mathcal{E}$ *if and only if* it is a necessary element of a set of sufficient conditions for $\mathcal{E}$ to happen. Let's see once again the voting example of the previous paragraph. "*Going to the*

NESS CAUSATION *v.* overdetermination

[7] The position that there are multiple types of CAUSATION is called *causal pluralism* and is a quite common position in philosophy of causation (*see e.g.* Russo 2023; Russo and Rihoux 2023; Braddon-Mitchell 2017; Lombrozo 2010; de Vreese 2006).

[8] Note that judging whether something is *reasonable*, *real & immediate* is once more a case of *evaluative judgements* that lead to all kinds of interpretational controversies (§1.1; *cf.* Stoyanova 2018, §§7.A-7.B)

beach" won with 8/10 votes. From all these 8 votes, any 6 of them are sufficient to go to the beach. I.e., every 6 of them is a *sufficient condition* to go to the beach. At the same time, for every such set of 6 votes, 1 vote is a *necessary* element for that set to be sufficient; without that 1 vote, the score would have been 5-5. In other words, that 1 vote is a *necessary element of a sufficient condition*. Ergo, that 1 vote is a *cause* according to the NESS test. Since the NESS test succeeds for every single vote of the 8 votes, all of them are NESS CAUSES. Therefore, every person that voted in favour of going to the beach is a NESS CAUSE of going to the beach.

ECtHR & overdetermination

The cases of the ECtHR are usually cases of overdetermination since there are multiple *sufficient* factors that contribute to a violation of the Convention. Note that for a causal justification of an ECtHR judgement, it is important to provide *all* the factors that contributed to the violation. Firstly, if a factor is not included, then we can not *foresee* that this factor will lead to violation of the law if it is repeated in the future (violation of the legitimacy requirement of *foreseeability*). Secondly, the more the factors that violate a right the more the harm. And inversely, the more the factors that mitigate the violation of a right, the lesser the harm. Determining the *severity* of the harm is decisive for deciding whether the Convention has been violated as well as which will be the *remedies* to the involved parties.

Let's see an actual example of CAUSAL OVERDETERMINATION in the ECtHR's case-law. I will use the *Perincek v. Switzerland (2015)* case that we have already encountered multiple times. The facts of the case are that the applicant denied the Armenian genocide at three public events, the state of Switzerland censored him *inter alia* by criminally convicting him, and the ECtHR judged that this censorship was a disproportional interference to the applicant's right to FREEDOM OF EXPRESSION (ARTICLE 10) compared to the harm that the applicant caused. The Court provided multiple justifications about its final judgement including the following two: p_1 :="*the statements cannot be regarded as affecting the dignity of the members of the Armenian community to the point of requiring a criminal-law response in Switzerland*", p_2 :="*the Swiss courts appear to have censured the applicant for voicing an opinion that diverged from the established ones in Switzerland and that the interference took the serious form of a criminal conviction*" (¶280). Both of p_1 & p_2 are *individually sufficient* to cause *harm* $\mathcal{E}$ to the applicant's RIGHT TO FREEDOM OF EXPRESSION violating ARTICLE 10: (α) a state interference disproportional to the harm caused by the applicant is a textbook case of Convention violation (*see e.g.* ECtHR Registry 2021, especially §III.B.3.b); (β) an interference to an opinion because it goes against the established views of the majority is also a textbook case of violating ARTICLE 10 (FREEDOM OF EXPRESSION) (*see e.g.* Press Service of the ECtHR 2020, p.1; ECtHR Registry, §XIV 2021).[9] Since both of them are sufficient to cause $\mathcal{E}$ by themselves, both of them *fail* the but-for test and ergo they can not be characterised as BUT-FOR CAUSES of $\mathcal{E}$. More precisely, even if the Swiss courts had not criminally convicted the applicant because his opinion was against the established opinions in Switzerland, they still did criminally convict him disproportionally for the harm he did to the dignity of the victims of the Armenian genocide. Similarly, even if Swiss courts had not criminally convicted the applicant due to the harm he did to the dignity of the victims, they still criminally convicted him because his opinion was against that of the established opinions in Switzerland. Consequently, based on the but-for test, neither $\mathcal{C}_1$:="*state interference disproportional to the harm to the dignity of the victims*" nor $\mathcal{C}_2$:="*severe censoring due to disagreement with the established views*" have *caused* the harm to the applicant's right to freedom of expression to the point of violating ARTICLE 10. One the other hand, based on the *NESS test*, both of $\mathcal{C}_1$ & $\mathcal{C}_2$ are NESS CAUSES causes of $\mathcal{E}$. More precisely, there are two sets of sufficient conditions for $\mathcal{E}$ to happen. One set is $\text{NESS}_1 = \{\mathcal{C}_1\}$ & the other one is $\text{NESS}_2 = \{\mathcal{C}_2\}$. Since $\mathcal{C}_1$ is the *only* element of NESS_1 then it is a *necessary* element for NESS_1 to make $\mathcal{E}$ true. The same argument holds for NESS_2.

One could counterargue that both $\mathcal{C}_1$ & $\mathcal{C}_2$ refer to the same act $\mathcal{C}'$, Switzerland criminally convicting the applicant, and ergo, they should be construed as a *single* cause $\mathcal{C}'$. If that is the case, the but-for test succeeds: *but-for* $\mathcal{C}'$, the harm $\mathcal{E}$ would not have happened. If we engage in this discussion, we enter the discipline of metaphysics asking questions of what *kind* of object can be a CAUSE and what *kind* of object can be an EFFECT. For the reader interested on the metaphysics of causation in law, they can have a look at Moore 2009 and its critical follow-up Gruyter 2013; *see also* Stapleton 2009. I will avoid entering into this discussion, a discussion that is more suited for a formal metaphysician, and I will show that even if one accepts this objection, there are still cases of NESS CAUSATION in the ECtHR's case-law. Let's assume that indeed what can be a cause is *acts* and not their properties (e.g., the property of an act *being disproportional to the harm caused by another act*). We saw that the *Perincek v. Switzerland (2015)* case, the Court argued that the statements of the applicant *did* harm the dignity of the victims, but they did not harm it enough to necessitate a criminal conviction. We also saw in §III.3.2.2 that one of the reasons why that harm was not sever enough was that the audience of the applicant's statements was that of three public events which the Court found to be rather limited. I.e., according to the Court, there is a correlation between the number of the audience $\mathcal{N}$ and the harm $\mathcal{E}'$ caused to the dignity

[9] There can always be exceptions in the case-law. Due to my lack of expertise and since identifying such exceptions falls outside the scope of the Thesis, I assume that there are no such exceptions. After all, we can always incorporate those exceptions by using *non-monotonic logic* or other methods used to deal with the paradox of MONOTONICITY (*cf.* §2.1.1, ¶3).

of the victims. Since we want the ALGOAI to predict *future* cases, we need to account for cases in which this audience $\mathcal{N}$ is *large enough* to cause severe enough harm $\mathcal{E}'$ to the dignity of the victims that would necessitate a criminal conviction or other type of state interference. If we have a case of at least two public events with such large number of audience, then *both* of those events are *sufficient* by themselves to cause the harm $\mathcal{E}'$. Ergo, neither of them is a BUT-FOR CAUSE, but both of them are NESS CAUSES (same arguments as in the previous paragraph). Another example is a state interfering to the publication of *multiple* publications each of which entails a sufficient condition to violate ARTICLE 10 (FREEDOM OF EXPRESSION) (*cf. Stomakhin v. Russia (2018); Dink v. Turkey (2010); Ibragim Ibragimov and others v. Russia (2018)*).

Concluding, my position is that both BUT-FOR & NESS CAUSATION are used as *explications* of FACTUAL CAUSATION, a position of CAUSAL PLURALISM (*fn.* 7). Moreover, Hart and Honoré's bifurcation of FACTUAL & LEGAL CAUSATION constitutes essentially the CAUSAL JUSTIFICATION used by the ECtHR to justify its jusdgements. Considering the foregoing, I continue to STEP II of explication: *formalising* SOURCE CONCEPTS introduced in STEP I.

IV.2 STEP II: formalising SOURCE CONCEPTS

As argued in §1, ¶4, contemporary formalisations of CAUSALITY, CAUSAL INFERENCE, CAUSAL RELATIONS, CAUSAL EXPLANATION, and the like have been heavily motivated by the use of those concepts in law and their application in jurisprudence with the standard reference point being Hart's & Honoré's 1959; 1985 SINE QUA NON CAUSATION. Two landmark explications stemming from the SINE QUA NON CAUSATION are Mackie's INUS[10] CAUSATION from the intersection of the disciplines of formal metaphysics & logic (1965; 1986) and Wright's NESS CAUSATION from the discipline of formal jurisprudence (1985; 1988; 2011). For a comparison among INUS, NESS, & SINE QUA NON CAUSATION *see* Wright 2011, §II; Baldwin 2003, §2. The birth of contemporary efforts to explicate CAUSATION *via* logic is traditionally considered to be John Stuart Mill's 1843 "*Systems of logic*" (Hart and Honoré 1985, §I.II; Pearl 2022, p.283). The turning point in the *philosophical* analysis of CAUSATION is traditionally considered to be the interpretation of CAUSATION by the Enlightenment philosophe David Hume (1739-1740; 2021; *cf.* Morris and Brown 2022, §5; Hart and Honoré 1985, §I.II), another Enlightenment effort to mechanistically interpret the world *via* reason (*cf.* §I.2.3; for a comparison between the interpretations of CAUSATION by the two trailblazer Enlightenment philosophes Immanuel Kant & David Hume *see* De Pierris and Friedman 2018). Due to the importance of causal inference in other disciplines, many contemporary conceptual re-engineerings of CAUSATION are performed with the intention for the re-engineered concept to be used in (semi-)formal models of diverse disciplinary practices. They are philosophical artifacts designed for non-philosophical (?) purposes. Two of the most impactful ones are Spirtes, Glymour, and Scheines 2000 (with the first edition being published in 1993) and Pearl 2009 (with the first edition being published in 2000). Spirtes, Glymour, and Scheines's work is used as a reference point for the engineering of (AI) algorithms that *extract causal relations* from sets of data, while Pearl's work is used as the gold standard of *formally modelling* those causal relations. A recent pioneering work that builds upon Pearl's *logical* explication of CAUSATION is Halpern 2016 (*see more* on its origins on *fn.* 13). Finally, other (semi-)formal explications that I consulted for the Thesis but I will not be use after all are Russo's 2009 explication of CAUSATION in *social sciences* and Canavotto's 2020 (§3.3) explication of the use of (NESS) CAUSATION in responsibility attribution by using a formal language of *action logic*.

(formal) explications of CAUSATION

Now that I introduced some landmark explications upon which I will base the formalisation of the SOURCE CONCEPTS, it is time to proceed with that formalisation. In §II.4.2.3, I proposed to follow a *hybrid* approach that forces specific formal structures to ALGOAI's output, those structures being formalisations of *legitimate judicial* reasoning. I further argued that this is should be done by identifying *formal subsumptive tests* that determine which justifications are *legitimate judicial justifications*. In our case, this means that I need to identify formal subsumptive tests so as to evaluate whether a justification is subsumed by the concept CAUSAL JUSTIFICATION that was explicated in §1.1.

To formalise CAUSAL JUSTIFICATION, we need to have a look at its potential FORMAL FORMS. To identify such a form I will take advantage of the fact that CAUSAL JUSTIFICATION is a HYPONYM of JUSTIFICATION. JUSTIFICATION is part of the material field of formal philosophy of explanation. More precisely, a justification can be construed as a rational argument that is used by humans to *epistemically access* true propositions. In the context of formal philosophy of explanation, JUSTIFICATION should be differentiated from EXPLANATION: albeit both have the same LOGICAL FORM, an explanation is used to explain what *is* the case in the world (*ordo essendi*), while a justification is used to justify *how* we come to *know* about what is the case in the world (*ordo cognoscendi*) (de Jong and Betti 2010, p.201). EXPLANATION is about the *ontology* of the world, while JUSTIFICATION is about our

identifying FORMAL FORM *via* RELATIONAL SIMILARITIES

[10] *Insufficient, but necessary part of an unnecessary but sufficient condition.*

epistemic access to the world. Take for instance again the example of Platonism and truths about mathematical objects (§II.3.1.2.1, *fn.* 23). Relations between Platonic objects can explain mathematical truths, but they can not explain our epistemic access to them. Hence, they can not be used as justifications. *Contra* to deductive mathematical proofs that can be used as justifications of why we know mathematical truths. Despite their differences, since in the context of formal philosophy of explanation EXPLANATION & JUSTIFICATION have the same LOGICAL FORM, I can identify JUSTIFICATION's FORMAL FORM by identifying EXPLANATION's LOGICAL FORM.

identifying FORMAL FORMS

What is the FORMAL FORM of EXPLANATION then? An instance of EXPLANATION is traditionally explicated as two RELATA, the EXPLANANDUM and the EXPLANANS, and a RELATION between the two *relata*, the EXPLANATORY RELATION. The semantical interpretation of the EXPLANATORY RELATION is that the EXPLANANS *explains* the EXPLANANDUM, but not the other way around (Woodward and Ross 2021, §2.5). That makes the EXPLANATORY RELATION between them to have the property of ASYMMETRY. Ergo, a NECESSARY INTENTIONAL RULE for the concept EXPLANATORY RELATION is the following: "*IF a relation is not asymmetric, THEN it is not subsumed by EXPLANATORY RELATION.*" Since this is also the LOGICAL FORM of JUSTIFICATION, one could propose to change the terms to describe this LOGICAL FORM (e.g., use "*justificatory relation*" instead of "*explanatory relation*") so as to know when we employ an explanation and when justification. Apart from THEORETICAL PARSIMONY, I reject this idea for two more reasons. Firstly, EXPLANATION & JUSTIFICATION are EXTENSIONALLY EQUIVALENT in the ECtHR's practice. The ECtHR is concerned only with the explanations that it can use to *epistemically access* true information. Secondly, the more the different terms conceptual engineers use, the more confusing their collaboration becomes undermining the EXPERT-ORIENTED UNDERSTANDABILITY.

Let's see an example of a justification used in the *Perincek v. Switzerland* and other historical negationism ECtHR cases (e.g., *Garaudy v. France*; *Lehideux & Isorni v. France*). The following is a SYNTACTICAL INTERPRETATION of that justification by using the LOGICAL FORM proposed in the previous paragraph:

a PARADIGMATIC EXAMPLE

$$\underbrace{\text{The applicant's denial of the genocide}}_{\text{EXPLANANS}} \quad \underbrace{\text{caused}}_{\text{EXPLANATORY RELATION}} \quad \underbrace{\text{harm to the dignity of the victims of the genocide.}}_{\text{EXPLANANDUM}}$$

Let's see whether the property of ASYMMETRY is present in this example. By knowing that the applicant denied the Armenian genocide, we can infer that the dignity of the victims of the said genocide was harmed. However, by knowing that the dignity of the victims of the Armenian genocide was harmed, we can not infer that the applicant denied the Armenian genocide since there can be many ways in which the dignity was harmed (e.g., the applicant said profanities against the victims). In other words, knowing that the EXPLANANS happened allows us to know that the EXPLANANDUM happened as well, but not the other way around. ence, the property of ASYMMETRY holds. In this example, the asymmetry of the explanatory relation is an EPISTEMIC ASYMMETRY since it is an asymmetry of what the user of this justification can *know*. Note that EPISTEMIC ASYMMETRY is crucial for the CORE PURPOSE of responsibility attribution in the legal tradition of the ECtHR when it comes to the RIGHT TO FREEDOM OF EXPRESSION (ARTICLE 10) as well as when it comes to identifying LEGAL CAUSATION: if the applicant can *foresee* that an act will lead to a specific harm and they proceed to perform that act, then the chances of being found responsible for the harm is raised. On the contrary, if the applicant could not have foreseen the outcome of their act, then the chances of being found responsible for the harm are lowered due to lack of FORESEEABILITY (ECtHR Registry 2021, §III.A.1; Stoyanova 2018, p.314 and §7.B). This also exhibits once more the importance of JUSTIFICATION to the REALISATION of the rule of law TRANSDI end of FORESEEABILITY.

REALISING TRANSDI ENDS

Since both RELATA of the foregoing example describe a state of the world (the state being that the applicant denied the Armenian genocide and that this utterance harmed the dignity of the victims), we can represent that state by employing the most common formal concept used to represent states of the world, that of PROPOSITION (*cf.* Alchourrón 2015, p.262; Hilpinen and McNamara 2021, pp.22-26):

more LOGICAL FORMS!

$$p := \text{"}\textit{The applicant denied the genocide.}\text{"}$$

$$q := \text{"}\textit{The dignity of the victims of the genocide was harmed.}\text{"}$$

Moreover, since the relation between p and q is asymmetric, we can use an asymmetric symbol between them to notate this asymmetry. I choose to use the symbol " $e=$". On its left, I will place the EXPLANANDUM and on its right the EXPLANANS: $q\ e{=}\ p$. I made this choice inspired by the symbol " $c=$" used by formal philosopher Nancy Cartwright (2002), a pioneer in formal philosophy of causation (*ibid.*; 1983; 2003; 2007; Cartwright and Efstathiou 2011), to notate the CAUSAL RELATION between the CAUSE (on the right of " $c=$", in the place of the EXPLANANS) and its EFFECT (on the left of " $c=$", in the place of the EXPLANANDUM). E.g., if $\mathcal{C}$ is a cause & $\mathcal{E}$ is its effect, their causal relation is represented as $\mathcal{E}\ c{=}\ \mathcal{C}$. I will use both " $e=$" & " $c=$" to notate explanatory relations & causal relations respectively. My intention for doing so is to show that in the

target system of concepts, there is a *relation* between CAUSAL RELATION & EXPLANATORY RELATION, that of HYPONYM-HYPERNYM respectively.[11]

Cartwright's symbol has another use that I adopt. Both p & q are propositional VARIABLES that can be assigned different TRUTH VALUES. That way, the same model can be used to represent different potential states of the *actual* world raising its UNIVERSALITY. E.g., it can represent states of the world in which p is TRUE (henceforth $\mathcal{T}$) as well as states of the world in which p is FALSE (henceforth $\mathcal{F}$). " $c=$ " is also used to notate that the values of the right side (CAUSE) influence the values of the left side (EFFECT). This is also the case for " $e=$" since we saw that knowing the truth values of the EXPLANANS can be used to acquire knowledge about the truth values of the EXPLANANDUM, but not the other way around. This type of equations between propositional variables is called in the literature of formal philosophy of causation *structural equations* (Pearl 2022, §.1.4.1; Peters, Janzing, and Schölkopf 2017, pp.9,83; Hall 2007; for the use of structural equations already from the 80s in the (formal philosophy of) econometrics *see* Aldrich 1989; for a different motivation for using structural equations *see fn.* 16).

A SEMANTICAL INTERPRETATION of structural equations that I adopt is that structural represent the *causal laws* in virtue of which the causal relations hold (Beckers 2021, pp.6212-6213; *cf.* Huber 2013). I.e., whenever the effect $\mathcal{E}$ is true due to the cause $\mathcal{C}$ being true, this happens in virtue of the causal law represented by $\mathcal{E}\ c{=}\ \mathcal{C}$. This is why CAUSAL LAWS are characterised as TRUTHMAKER. Both LAWS and CAUSATION are needed to have an understanding of the world, and subsequently, to *justify* true propositions about that world:

> "*'Cause' and 'law' are perhaps two of the most* ***important and fundamental concepts*** *that human beings deploy in their attempts to* ***understand and intervene*** *in their environment, both in everyday life and in their scientific endeavors. When we want to explain why an event occurred, we seek out its causes. When we act, we do so because we believe the action will have certain effects. When we want to know why our environment and our fellow human beings behave in regular, predictable ways, we look for the laws that govern that behavior.*" Beebee 2015, p.266, emphasis added

SEMANTICAL INTERPRETATIONS

I adopt the SEMANTICAL INTERPRETATION of structural equations as causal laws since it allows me to make a further differentiation between JUSTIFICATION & CAUSAL JUSTIFICATION. More precisely, *causal relations* are true *in virtue of* CAUSAL LAWS, while *explanatory relations* can be true *in virtue of* both CAUSAL & other NON-CAUSAL type of "*laws*" like regularities premissed on statistical correlations like those of connectionist AI that we saw in §I.3.2.1.1, ¶11 (Kistler 2013; *cf.* Beckers 2021, pp.6212-6213). Ergo, we can GENERALISE the SEMANTICAL INTERPRETATION of structural equations from representing causal laws to representing a superset of causal laws (HYPONYM-HYPERNYM RELATION). That being said, I do not claim that every type of LAW can be represented by structural equations. For instance, we saw that both structural equations & explanatory relations are *asymmetric*. Ergo, *non*-asymmetric laws, assuming there are such laws, can not be represented by structural equations.

Let's sum up everything explicated so far using the dominating discipline's system of concepts. CAUSAL JUSTIFICATION is a particular case of JUSTIFICATION that the ECtHR uses in its judgements. The FORMAL FORM of both is that of STRUCTURAL EQUATIONS which represent (CAUSAL) LAWS, and subsequently the METAPHYSICAL & EPISTEMIC COMMITMENTS of these explications are those that result from accepting the existence of (CAUSAL) laws, whatever one construes as such.[12]. Considering the above, what we need at this point is a *subsumptive test* (i.e., INTENTIONAL RULE OF USE) that makes use of that FORMAL FORM so as to determine when a justification is a *CAUSAL* JUSTIFICATION. Taking into consideration the explication of CAUSAL JUSTIFICATION in §1, due to the ECtHR's CAUSAL PLURALISM, we need *two* subsumptive tests: (α) the BUT-FOR CAUSATION subsumptive test; "*A particular c is a BUT-FOR CAUSE of a BUT-FOR EFFECT ε IF AND ONLY IF c is a necessary element for ε to be true.*" (β) the NESS CAUSATION subsumptive test: "*A particular c is a NESS CAUSE of a NESS EFFECT ε IF AND ONLY IF c is a necessary element of a sufficient condition for ε to be true.*". The question is now how can we *formalise* these tests by using CAUSAL JUSTIFICATION's FORMAL FORM. This is the question I answer in the next subsection. Before doing so, not that henceforth, by "*explanatory model*", I will be referring to the explication of JUSTIFICATION, and by "*causal model*" I will be referring to the explication of CAUSAL JUSTIFICATION.

[11] For an alternative formal philosophical inquiry on the (SEMI-)LOGICAL FORM of CAUSAL RELATION *see* Davidson's 1967 seminal as well as Widerker's 1985 overview of the Davidsonian approach.

[12] For a classical formal metaphysical account of CAUSAL LAWS *see* Armstrong's seminal "*What is a law of nature?*" (1983; *cf.* Heathcote and Armstrong 1991; Armstrong 1997, §§15-16). For a renowned endeavour in the literature of formal philosophy of explanation to justify humans' epistemic access to causal relations without appealing to causal laws or any other metaphysically binding arguments *see* Woodward 2003, 2007, 2015. I do not construe Woodward 2003's construal of CAUSATION as a Carnapian *explication* since Woodward himself (2007) denied that his work constitutes any sort of *conceptual* analysis, let alone of *conceptual* engineering (Strevens 2008, p.184; *cf.* Diakite 2016, footnote 1).

IV.2.1 Two subsumptive test

IV.2.1.1 The semantics

A recent proposal for a formal logical explication of BUT-FOR & NESS subsumptive tests is that of Beckers 2021 which is based on the semantics of the seminal *HP* logical explication of CAUSATION[13] (*cf.* §2, ¶1; *see also* Beckers and Vennekens 2017, 2018 for previous works of the author that paved the way for their 2021 paper). In what follows, I will show step by step how Beckers's explication seems[14] to be indeed an *adequate* explication of BUT-FOR & NESS CAUSATION. For a different approach on using *structural equations* to explicate NESS CAUSATION *see* Baldwin 2003.[15]

To introduce the subsumptive tests, I first need to formalise more aspects of CAUSAL JUSTIFICATION. Unless specified otherwise, the formal semantics I am using are those of Halpern 2016 so as to be *on par* with Beckers's semantics (*cf. fn.* 13). First things first, we saw that structural equations are composed by propositional variables that can represent potential states of the world. For reasons of SEMANTICAL SIMPLICITY (or REPRESENTATIONAL STUPIDITY if you prefer), we are interested in using propositional variables to represent *only a few* aspects of the world. The choice of which of those aspects are important will be done by legal scientists (e.g., by using the criteria that discern LEGAL CAUSATION from mere FACTUAL CAUSATION (§1, ¶¶4-5)). It is common practice in causal modeling to call the variables that represent those specific aspects of the world as *endogenous variables* (they are "*endo*" to our model) and notate their set as $\mathcal{V}$. At the same time, for reasons of RELATIONAL SIMILARITY, we should still account for the effect that the rest of the world has on those specific variables. Hence, for every endogenous variable $X \in \mathcal{V}$, we should introduce at least one variable U_X which represents that influence. It is common practice to call such variables *exogenous* (they are "*exo*" of our model) and notate the set of all such variables as $\mathcal{U}$. It is also common practice to use only *one* exogenous variable U_X *per* endogenous variable X which can again be justified for reasons of SEMANTICAL SIMPLICITY, as well as for reasons of MINIMAL COMPLEXITY, SYNTACTICAL SIMPLICITY and UNDERSTANDABILITY. If we include exogenous variables in the model of §III.1, Figure 1, we get the following model

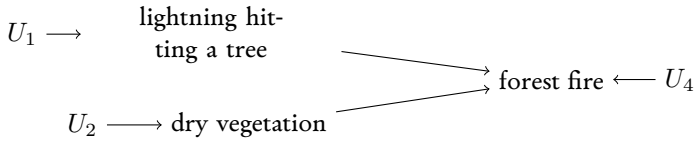


Figure 1: This is a toy model of causal inference. It is a variation of Figure 2.1 (b) in Halpern 2016, p.16.

Are those variables enough to describe the world? Almost! The inclusion of variables entails the inclusion of the *range of values* that those variables can take. Except from certain *suis generis* logics, most logics include in their range of values the truth values TRUE ($\mathcal{T}$) and FALSE ($\mathcal{F}$). For reasons of PARSIMONY, it seems more preferable to include only those two values in our model. However, it is common practice in logic to include more truth values so as to accomodate LOGICAL PARADOXES. For instance, it is common practice in logic-based legal AI to use *non-monotonic* logic so as to accommodate the NON-MOTONICITY of judicial reasoning (*see e.g.* Gordon 1988; Sartor 2012; Rigoni 2014; Iatrou 2022b, 2022a). Usually, non-monotonic logics have a third truth value of DEFAULT FALSEHOOD ($\mathcal{DF}$) (Lukaszewicz 1990, §5; Wan, Kifer, and Grosof 2015; Strasser and Antonelli 2019, §3.3): everything that is not made explicitly true in our model is false *by default*. The assumption that only the world represented explicitly by the model is true is the so-called *closed world assumption* (Lukaszewicz 1990, p.105; Strasser and Antonelli 2019, §3.1). DEFAULT FALSEHOOD is a different truth value from explicitly declaring something FALSE in the model, and subsequently, it requires a different logical treatment. A PARADIGMATIC use of DEFAULT FALSEHOOD in legal AI is its ability to model the principle of the PRESUMPTION OF INNOCENCE according to which every accusation for breaking the law is *by default* FALSE with the accuser having the burden of proving before a court that the law has been violated (Cabalar, Fandinno, and Fink 2014; Iatrou 2022b; *cf.* Robertson 2009, p.1009,§C.2). PRESUMPTION OF INNOCENCE is particularly important for the legitimacy of political orders falling under the ECtHR's jurisdiction since it is a HUMAN RIGHT protected by ARTICLE 6 (RIGHT

accommodating LOGICAL PARADOXES

[13] "*HP*" stands for the intials of (Joseph Y.) *H*alpern & (Judea) *P*earl who are the instigators of this explication. The origin of the explication is Halpern and Pearl 2005 and the updated explication that Beckers 2021 uses is the one in Halpern 2016. Precursors of the HP definition are Galles and Pearl 1997; Pearl 2009; Halpern 2000.

[14] I write "*seems*" since as already argued this chapter is more of a toy example of how the first phase of CAUSAL JUSTIFICATION's explication should look like and not a full-fledged account (*see e.g.* the introductions of CHAPTERS III & IV).

[15] Note that Baldwin's proposal was published before Wright's last defense & disambiguation of NESS CAUSATION in 2011, while Beckers 2021 was published 10 years after, with Beckers citing mostly Wright's 2011 paper and not their previous ones.

TO A FAIR TRIAL), ¶2 (*cf.* ECtHR's Registry 2022a, §VI.A; ECtHR's Registry 2022b, §1.I.E.4).

Another challenge regarding the *range of values* of the model's variables is whether we will assign *one* value *per* propositional variable. *Prima facie*, for reasons of LOGICAL CONSISTENCY that should not be the case. However, we can once more opt to reject LOGICAL CONSISTENCY so as to accommodate LOGICAL PARADOXES. For instance, in the cases of DEFAULT FALSEHOOD that we just saw, DEFAULT FALSEHOOD can be updated by another truth value allowing the same proposition to take two TRUTH values in the same model, albeit not simultaneously. Take the example of the PRESUMPTION OF INNOCENCE. We saw that proposition that the defendant is legally responsible for a harm should be considered FALSE by default. However, if the facts of a case prove that the defendant is indeed legally responsible for that harm, then that proposition should change its truth value to TRUE.

more LOGICAL PARADOXES

Whatever one chooses regarding truth values, and what I choose is to use classical two-valued logic since there is no reason to make this chapter more complicated, for a specific state of the world, we have a function $\mathcal{R}$ that corresponds every variable $X \in \mathcal{U} \cup \mathcal{V}$ to a specific set of values $\mathcal{R}(X)$. Ergo, the triplet $(\mathcal{U}, \mathcal{V}, \mathcal{R})$ that I will call *signature*, signifies a specific part of the world, the part represented by the variables $\mathcal{U} \cup \mathcal{V}$. For reasons of SYNTACTICAL SIMPLICITY & EXPERT-ORIENTED UNDERSTANDABILITY, it is common practice to annotate a collection of variables $X_i \in \mathcal{U} \cup \mathcal{V}$ as a vector $\bar{X} = (X_1, X_2, ..., X_n)$ with image $\mathcal{R}(\bar{X}) = (\mathcal{R}(X_1), \mathcal{R}(X_2), ..., \mathcal{R}(X_n))$.[16] If a variable X is equal to a value x, I will symbolise it as $\bar{X} = \bar{x}$ and I will call it an *atomic formula*. I will use the same term for $\bar{X} = \bar{x}$. Finally, a *complex* formula will be a formula composed of atomic formulae by using logical operators. E.g., $\varphi = \bar{X_1} = \bar{x_1} \wedge \bar{X_2} = \bar{x_2}$. In this Thesis, I will be using the traditional logical operators $(\wedge, \vee, \rightarrow, \neg)$, albeit one can use different operators to accommodate for LOGICAL PARADOXES. E.g., in the example of non-monotonic logics with DEFAULT FALSEHOOD, there is usually an extra logical operator which signifies when a proposition is FALSE by default, the DEFAULT NEGATION operator.

In §2, ¶¶7-8, we saw that apart from variables and their values, an explanatory model includes *structural equations*. In other words, we have another function $\mathcal{F}$ that assigns to each variable $\bar{X}$ a structural equation $f_{\bar{X}}$ whose arguments are the variable $\mathcal{V} \cup \mathcal{U}$:

$$X \; e{=} \; \mathcal{F}(X), \text{ where } \mathcal{F}(X) := f_{\bar{X}}(\mathcal{V} \cup \mathcal{U})$$

For instance, the explanatory model of Figure 1 has the following structural equations:

$$\begin{aligned} FF \; e{=} \; & f_{FF}(LT, DV, U_1, U_2, U_3) = LT \wedge DG \wedge U_3 \\ DV \; e{=} \; & f_{DV}(LT, FF, U_1, U_2, U_3) = U_2 \\ LT \; e{=} \; & f_{LT}(FF, DV, U_1, U_2, U_3) = U_1 \end{aligned}$$

where FF :="*There is a forest fire.*", LT :="*A lightning hits a tree.*", and DV :="*The vegetation is dry.*". Note that it is common practice to ignore the exogenous variables in a structural equation unless they are the *only* variables of the structural equation. I.e., $f_{FF}(LT, DV, U_1, U_2, U_3)$ becomes simply $LT \wedge DG$ while f_{DV} and f_{LT} remain as is. Furthermore, exogenous variables are usually not assigned structural equations. They are assigned directly a value u from $\mathcal{R}(\mathcal{U})$ since we only care about which *are* the background conditions and not *how* they come to be. The vector $\bar{u}$ of the exogenous variables' values $(\bar{U} = \bar{u})$ is called the *context* of the model. I will call the 2-tuple of a signature $\mathcal{S}$ and the function $\mathcal{F}$ that assigns to $\mathcal{S}$'s variables structural equations an *explanatory model of the world*. I will annotate any such model of the world as $\mathcal{M} = \langle \mathcal{S}, \mathcal{F} \rangle$.

It is worth noting that we can use the function $\mathcal{F}$ to accommodate LOGICAL PARADOXES in our model. E.g., we can model AMBIGUITY, i.e. OVERDETERMINATION, by assigning different structural equations to the same variable, albeit then $\mathcal{F}$ will not be a function. E.g., we can engineer a model that has both structural equations $FF \; e{=} \; LT$ and $FF \; e{=} \; \neg LT \wedge DV$. We can also model VAGUENESS, i.e. UNDERDETERMINATION, by assigning to a variable a piecewise structural equation with underdetermined conditions. E.g., in Figure 1, we assign to FF a structural equation for U_3 being TRUE, but we do not assign any structural equation for U_3 being FALSE (*cf.* Fine 1975, p.266).

even more LOGICAL PARADOXES

For a specific part of the world (i.e., a specific signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$), to determine which variables are causes of which effects, we should inquire how the endogenous variables interact with each other while disregarding other information that can influence their behaviour. I.e., we are interested in how the endogenous variables interact while keeping the rest of the world fixed; we are interested in how the endogenous variables interact with each other in specific context $\bar{u}$. The 2-tuple $(\mathcal{M}, \bar{u})$ will be called *explanatory setting*. $(\mathcal{M}, \bar{u}) \models X = x$ or simply $\mathcal{M}, \bar{u} \models X = x$ annotates that the atomic formula $X = x$ is true in the explanatory setting

[16] Using *vectors* as an alternative notation is also motivated by the extended use of causal inference in the disciplines of statistics, econometrics, & probability theory. In those disciplines, the use of vectors is common practice. A similar motivation holds for the choice to use *structural equations* to model causal relations (Halpern 2016, pp.6-7,12,21-22). Thus, these choice of notation corroborate further the UNI-

$(\mathcal{M}, \bar{u})$. Similarly, $(\mathcal{M}, \bar{u}) \models \bar{X} = \bar{x}$ or simply $\mathcal{M}, \bar{u} \models \bar{X} = \bar{x}$ annotates that the atomic formulae $X_i = x_i$ are true for every $1 \leq i \leq dim(\bar{X})$. Finally, regarding complex formulae, their truth is defined *per* usual. E.g., for the complex formula $\varphi = \bar{X_1} = \bar{x_1} \wedge \bar{X_2} = \bar{x_2}$ we have that $\mathcal{M}, \bar{u} \models \varphi$ *iff* $(\mathcal{M}, \bar{u} \models \bar{X_1} = \bar{x_1}$ and $\mathcal{M}, \bar{u} \models \bar{X_2} = \bar{x_2})$.

We are one step before finishing the introduction of the semantics that will be used to explicate the NESS & BUT-FOR subsumptive tests. We still need to introduce another FRUITFUL CONCEPT from the formal philosophy of causation discipline, that of INTERVENTION. ore precisely, to inquire how a variable X influences a variable Y, it is useful to change the value of X and see how it influences Y. However, if we want to inquire *only* the influence of X to Y and not that of other variables, we should change X while keeping the rest variables as they are. In other words, an INTERVENTION is when the structural equation of an endogenous variable X is substituted by a specific value x, while the rest of the structural equations stay as they are. Let's GLUE INTERVENTION to the rest of our target system. Assume a explanatory model $\mathcal{M} := \langle \mathcal{S}, \mathcal{F} \rangle$. Then, the operation of INTERVENTION on a variable $\bar{X}$ is symbolised as $\bar{X} \leftarrow \bar{x}$ and the new explanatory model is symbolised as $\mathcal{M}^{\bar{X} \leftarrow \bar{x}} := \langle \mathcal{S}, \mathcal{F}^{\bar{X} \leftarrow \bar{x}} \rangle$, where the function $\mathcal{F}^{\bar{X} \leftarrow \bar{x}}$ is defined as follows:

more FRUITFUL CONCEPTS

$$\mathcal{F}^{\bar{X} \leftarrow \bar{x}} = \begin{cases} \mathcal{F}_Y, & \forall Y \text{ s.t. } Y \neq X_i, \text{ where } 1 \leq i \leq dim(X) \\ x_i, & \text{otherwise} \end{cases}$$

In other words, $\mathcal{F}^{\bar{\mathbf{X}} \leftarrow \bar{\mathbf{x}}}$ is the same as $\mathcal{F}$ for all the variables for which we do not intervene, but it becomes x_i for all interventions $X_i \leftarrow x_i$.

Assume now that φ represents the facts of a case bought before a court and we want to see what would have happened if certain aspects of those facts had been different. E.g., *but-for* the defendant not shooting, would the victim have died? What we have to do is to change the values of th variables that represent the aspects of reality that we want to change while keep the rest aspects of reality the same. I.e., we need to perform an intervention $\bar{X} \leftarrow \bar{x}$. Those changes in φ will be symbolised as $[\bar{X} \leftarrow \bar{x}]\varphi$. For a specific explanatory setting $(\mathcal{M}, \bar{u}) := \langle \mathcal{S}, \bar{u} \rangle$, we have that $\mathcal{M}, \bar{u} \models [\bar{X} \leftarrow \bar{x}]\varphi$ *if and only if*[17] $\mathcal{M}^{\bar{X} \leftarrow \bar{x}}, \bar{u} \models \varphi$. Formulae of the form $[\bar{X} \leftarrow \bar{x}]\varphi$ will be called *explanatory formulae (over* $\mathcal{S}$*).*

Now we are ready to introduce the NESS & BUT-FOR subsumptive tests. I will first introduce Beckers's 2021 BUT-FOR subsumptive test, I will show why it fails to explicate cases of CAUSAL OVERDETERMINATION, and then, I will show how Beckers's NESS test, which is a modification of the BUT-FOR test allowing a straightforward comparison between the two, succeeds.

Before introducing the BUT-FOR test, I would like to respond to a reasonable objection to the construal of Beckers's 2021 formalisation of BUT-FOR & NESS CAUSATION as *explications*. When Beckers introduces those formalisations, they introduce them as *definitions* of BUT-FOR & NESS CAUSATION. While DEFINITION is a method of conceptual (re-)engineering, it is a *different* method from EXPLICATION (Brun 2016, §4.1) and ergo Beckers's formalisations seem ill-fit to serve as explications. This is another case of the same term being used to denote different concepts (§III.2.2.1.1). DEFINITION as a method of conceptual (re-)engineering a concept $\mathcal{C}$ of a system of concepts $\mathcal{S}$ refers to the use of concepts from the same system $\mathcal{S}$ to define $\mathcal{C}$ (Brun 2016, §4.1). What Beckers does though is to define pre-formal intuitive concepts of CAUSATION using concepts from a *different* system of concepts $\mathcal{S}'$, that of Halpern 2016 ($\mathcal{S} \neq \mathcal{S}'$). Ergo, Beckers's definitions can serve as explications since they belong to a different system of concepts of the *explicanda*. Definitions that use concepts from the target system of concepts are legitimate part of the process of explicating the *explicandum* (Brun 2016, §4.1).

DEFINITION ≠ DEFINITION

IV.2.1.2 The tests

IV.2.1.2.1 The BUT-FOR subsumptive test

As argued in §III.3.2.1, subsumptive tests can be construed as INTENTIONAL RULES OF USE. In what follows, I introduce an intentional rule of use that Beckers 2021 uses to decide whether a particular is a BUT-FOR CASE:

INTENTIONAL RULE (but-for test).
Let $(\mathcal{M}, \bar{u})$ an explanatory setting, $\mathcal{C}$ & $\mathcal{E}$ endogenous variables that represent an *explanans* & an *explanandum* respectively, and c & ε values of $\mathcal{C}$ and $\mathcal{E}$ respectively (i.e., $c \in \mathcal{R}(\mathcal{C})$ and $e \in \mathcal{R}(\mathcal{E})$). $\mathcal{C} = c$ is a *but-for cause* of $E = \varepsilon$ IF AND ONLY IF the following two conditions hold:

fiCATION with those disciplines, as well as the EXPERT-ORIENTED UNDERSTANDABILITY with regards to statisticians and the like (*cf. ibid.*).

[17] Beckers uses "*if*" and not "*if and only if*" (2021, p.6211). My interpretation is that despite saying "*if*", they mean "*if and only if*". This is corroborated by looking at Halpern 2016, p.21 where Halpern *explicitly* uses "*if and only if*". Even if Beckers wants to differentiate his approach from Halpern, I will follow Halpern's approach. The same holds for more FRUITFUL CONCEPTS that will be introduced from Beckers's paper later on.

- but-for condition 1: $\mathcal{M}, \bar{u} \models \mathcal{C} = c \wedge \mathcal{E} = \varepsilon$;
- but-for condition 2: $\exists c' \in \mathcal{R}(\mathcal{C})$ s.t. $c' \neq c$ & $\mathcal{M}, \bar{u} \models [\mathcal{C} \leftarrow c']\neg\mathcal{E} = \varepsilon$

$\mathcal{E} = \varepsilon$ is the *but-for effect* of $\mathcal{C}$ and the explanatory relation between the two is a BUT-FOR CAUSAL RELATION. but-for condition 1 represents that both $\mathcal{C} = c$ & $\mathcal{E} = \varepsilon$ describe a state of the world that *actually* happened *contra* other accounts of causation which concern *hypothetical* states of the world. This is of particular importance for responsibility attribution in *law* since judicial authorities should hold a defendant responsible based on acts that they have *actually* performed (Wright 2011, pp.286-287). but-for condition 2 represents that had the the value of $\mathcal{C}$ been different (c' instead of c), then $\mathcal{E} = \varepsilon$ would not have happened.

evaluating formalisations

Now we need to *evaluate* whether this formal conception of BUT-FOR CAUSATION captures its pre-formal conception. One way to do so is to see whether its pre-formal intentional rules are satisfied by this formalisation. This is more or less the approach that Beckers 2021 follows (p.6212). I would like to take a different approach though to which Beckers is rather condescending, but one which satisfies the SIMILARITY requirement of COHERENCE (§§III.3.2.1,III.3.3) and which helps to lift Benacerraf's curse (§I.1.1.1). I would like to test whether the but-for subsumption test identifies as but-for causes PARADIGMATIC EXAMPLES of but-for causes. Each ALGOAI engineering discipline has its own PARADIGMATIC EXAMPLES. E.g., legal scientists can test whether it identifies PARADIGMATIC EXAMPLES of CAUSAL JUSTIFICATION in the ECtHR's practice and formal philosophers & logicians can use PARADIGMATIC EXAMPLES from the literature of formal philosophy of causality. It is also advisable for AI engineers to evaluate whether PARADIGMATIC EXAMPLES of causal XAI can accommodate the proposed but-for test and under which conditions they do so (e.g., is the COMPLEXITY of the but-for test low enough to be used in logic-based discovery of causal structures like that of Hyttinen, Eberhardt, and Järvisalo 2014, a model that can not withstand moderate COMPLEXITY?). Finally, note that this is once more a case where an explication requirement is also an *evaluation* criterion (§III.3.2, ¶1): according to the COHERENCE requirement, we want to engineer an explication that satisfies PARADIGMATIC EXAMPLES and we can evaluate whether an explication realises COHERENCE adequately by testing whether it satisfies those examples.

positive v. negative paradigmatic examples

Lets put the but-for test to the test! In the PARADIGMATIC EXAMPLES that follow, I provide that *facts of a case* and which is the *natural cause*. What we want is the but-for test to satisfy both POSITIVE & NEGATIVE PARADIGMATIC EXAMPLES (*cf.* §III.3.3, ¶5).

EVALUATION: POSITIVE PARADIGMATIC EXAMPLE.
- **Facts of the case:** Patrick throws a rock at a bottle. The bottle breaks
- **Natural cause:** Patrick throwing the rock is the natural cause of the bottle breaking.

$$\text{Patrick} \longrightarrow \text{Bottle}$$

$$Bottle\ e{=}\ Patrick \wedge U_2$$

$$Patrick\ e{=}\ U_1$$

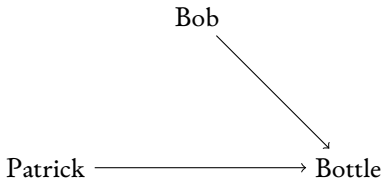
According to the facts of the case, we have that $(U_1, U_2, Patrick, Bottle) = (\mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T})$. We want to test whether $Patrick = \mathcal{T}$ is the but-for cause of $Bottle = \mathcal{T}$. In other words, $\mathcal{C} := Patrick$, $\mathcal{E} := Bottle$, $c =: \mathcal{T}$, and $\varepsilon := \mathcal{T}$. Do the two *but-for conditions* hold?

- but-for condition 1: We want to test whether $\mathcal{M}, \bar{u} \models \mathcal{C} = c \wedge \mathcal{E} = \varepsilon$. I.e., whether $\mathcal{M}, (\mathcal{T}, \mathcal{T}) \models Patrick = \mathcal{T} \wedge Bottle = \mathcal{T}$. And it holds.;
- but-for condition 2: We want to test whether $\exists c' \in \mathcal{R}(\mathcal{C})$ s.t. ($c' \neq c$ and $\mathcal{M}, \bar{u} \models [\mathcal{C} \leftarrow c']\neg\mathcal{E} = \varepsilon$). Indeed, for $c' = \mathcal{F} \in \mathcal{R}(Patrick) = \{\mathcal{T}, \mathcal{F}\}$, we have that $\mathcal{M}^{Patrick \leftarrow c'}, (\mathcal{T}, \mathcal{T}) \models \neg Bottle = \mathcal{T}$.

The but-for test is successful! Ergo, the explanatory relation is a BUT-FOR CAUSAL RELATION. Both $Bottle\ e{=}\ Partick$ & $Bottle\ c{=}\ Partick$ hold!

Let's see now how the BUT-FOR test fails in a case of overdetermination. For reasons of *EXPERT*-ORIENTED UNDERSTANDABILITY, I will use the same setting as the previous evaluative tests. Hence, instead of a voting example that we saw in §1.1, ¶5, I will use a Patrick-throws-a-rock example.

EVALUATION: NEGATIVE PARADIGMATIC EXAMPLE (overdetermination).
- **Facts of the case:** Both Patrick and Bob throw a rock & both rocks hit simultaneously the bottle breaking it.
- **Natural causes:** Both Patrick and Bob throwing a rock are natural causes of the bottle breaking.



$$Bottle\ e{=}\ (Patrick \vee Bob) \wedge U_3$$
$$Bob\ e{=}\ U_2$$
$$Patrick\ e{=}\ U_1$$

According to the facts of the case $(U_1, U_2, U_3, Patrick, Bob, Bottle) = (\mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T})$. We want to test whether both $Bob = \mathcal{T}$ and $Patrick = \mathcal{T}$ are *not* but-for causes of $Bottle = \mathcal{T}$. Due to symmetry, without loss of generalisation, I will only test whether $Patrick = \mathcal{T}$ is a but-for cause. Consequently, for this specific test, we have that $\mathcal{C} =: Patrick$, $\mathcal{E} := Bottle$, $c = \mathcal{T}$, and $\varepsilon = \mathcal{T}$. Do the two but-for conditions hold?

- but-for condition 1: We want to test whether $\mathcal{M}, \bar{u} \models \mathcal{C} = c \wedge \mathcal{E} = \varepsilon$. I.e., whether $\mathcal{M}, (\mathcal{T}, \mathcal{T}, \mathcal{T}) \models Patrick = \mathcal{T} \wedge Bottle = \mathcal{T}$. And it holds.;
- but-for condition 2: We want to test whether $\exists c' \in \mathcal{R}(\mathcal{C})$ s.t. ($c' \neq c$ and $\mathcal{M}, \bar{u} \models [\mathcal{C} \leftarrow c']\neg\mathcal{E} = \varepsilon$). However, for every possible $c' \in \mathcal{R}(Patrick) = \{\mathcal{T}, \mathcal{F}\}$, we have that $\mathcal{M}^{Patrick \leftarrow c'}, (\mathcal{T}, \mathcal{T}, \mathcal{T}) \models \neg Bottle = \mathcal{T}$ is false.

The but-for test fails! Hence, $Patrick = \mathcal{T}$ is *not* a but-for cause of $Bottle = \mathcal{T}$. Ergo, the explanatory relation $Bottle\ e{=}\ Partick$ is *not* a BUT-FOR CAUSAL RELATION. Due to symmetry, the same holds for $Bottle\ e{=}\ Bob$. □

Can Beckers's 2021 explication of the NESS test remedy but-for test's inadequacy?

IV.2.1.2.2 The NESS subsumptive test

To introduce the NESS test, Beckers 2021 firstly introduces an explication of SUFFICIENCY: an *explanans* is subsumed by the concept SUFFICIENT EXPLANANS IF AND ONLY IF whenever we know that the *explanans* happened, we know that the *explanandum* will *always* follow *no matter* what is going on to the rest of the world .[18] Beckers's formalisation of SUFFICIENT EXPLANANS is the following (2021, p.6213):

INTENTIONAL RULE (SUFFICIENT EXPLANANS' subsumption test I). For a collection of endogenous variables $\bar{X}$ and a different endogenous variable $\mathcal{E}$, we say that $\bar{X} = \bar{x}$ is a SUFFICIENT EXPLANANS for $\mathcal{E} = \varepsilon$ IF AND ONLY IF $f_{\mathcal{E}}(\bar{x}, \bar{z}) = \varepsilon$ for all possible values $\bar{z} \in \bar{Z}$. In other words, if we set the values of $\bar{X}$ to $\bar{x}$, the structural equation $f_{\mathcal{E}}$ of $\mathcal{E}$ outputs ε no matter what values the rest of the variables take. □

Beckers provides a second subsumption test for SUFFICIENT EXPLANANS. Albeit it is a test with lower EXPERT-ORIENTED UNDERSTANDABILITY, it is a quite FRUITFUL test since *contra* to the first test, it will be used to explicate the NESS test.

INTENTIONAL RULE (SUFFICIENT EXPLANANS' subsumption test II). For a collection of endogenous variables $\bar{X}$ and a different endogenous variable $\mathcal{E}$, $\bar{X} = \bar{x}$ is a SUFFICIENT EXPLANANS for $\mathcal{E} = \varepsilon$ with regards to a causal setting $(\mathcal{M}, \bar{u})$ IF AND ONLY IF for every possible value $\bar{z}$ that the rest of endogenous variables $\bar{Z}$ can take (i.e., all the endogenous variables Z_i that are neither X_j nor $\mathcal{E}$), it holds that $\mathcal{M}, \bar{u} \models [\bar{X} \leftarrow \bar{x}, \bar{Z} \leftarrow \bar{z}]\mathcal{E} = \varepsilon$. □

Now we are ready to introduce Beckers's NESS test:[19]

INTENTIONAL RULE (NESS CAUSATION subsumption test). Let $(\mathcal{M}, \bar{u})$ an explanatory setting, $\mathcal{C}$ & $\mathcal{E}$ endogenous variables that represent an *explanans* & an *explanandum* respectively, and c & ε values of $\mathcal{C}$ and $\mathcal{E}$ respectively (i.e., $c \in \mathcal{R}(\mathcal{C})$ and $e \in \mathcal{R}(\mathcal{E})$). $\mathcal{C} = c$ is a NESS CAUSE of $E = \varepsilon$ IF AND ONLY IF the following two conditions hold:

[18] Henceforth, for reasons of convenience, I will be introducing material from the cited sources expressed directly in the systems of concepts laid out until the previous subsection. I have already exhibited multiple times how to make the translations so henceforth I want to elaborate on the rest aspects of explications.

[19] In reality, the subsumptive test that follows can be used to identify *only direct* NESS causes. I.e., if there is a chain of causes between a NESS cause $\mathcal{C}$ and a NESS effect $\mathcal{E}$ (e.g., $\mathcal{C}$ first causes $\mathcal{C}_1$ which then causes $\mathcal{C}_2$ which then causes $\mathcal{E}$), then the following NESS substantive

- NESS condition 1: $\mathcal{M}, \bar{u} \models \mathcal{C} = c \wedge \mathcal{E} = \varepsilon$;

- NESS condition 2: there exists a $\bar{W} = \bar{w}$ s.t.

  2.α $(\mathcal{C}, W_1, W_2, ..., W_n) = (c, w_1, w_2, ..., w_n)$ is a sufficient *explanans* for $\mathcal{E} = \varepsilon$ w.r.t. $(\mathcal{M}, \bar{u})$;

  From the NESS subsumptive test II, NESS condition 2.α is equivalent to the following: for every value $\bar{z}$ that the rest of the endogenous variables can take (i.e., the endogenous variables that are not $\mathcal{C}$, W_i, $\mathcal{E}$) we have that $\mathcal{M}, \bar{u} \models [\vec{\mathcal{C}} \leftarrow \bar{c}, \bar{W} \leftarrow \bar{w}, \bar{Z} \leftarrow \bar{z}]\mathcal{E} = \varepsilon$;

  2.β $\bar{W} = \bar{w}$ is *not* a sufficient *explanans* for $\mathcal{E} = \varepsilon$ w.r.t. $(\mathcal{M}, \bar{u})$.

  From the NESS subsumptive test II, NESS condition 2.β is equivalent to the following: there exists a value $\bar{z}$ for the rest of the endogenous variables (i.e., the endogenous variables that are not W_i, $\mathcal{E}$) s.t. $\mathcal{M}, \bar{u} \models [\bar{W} \leftarrow \bar{w}, \bar{Z} \leftarrow \bar{z}]\neg\mathcal{E} = \varepsilon$.

  $E = \varepsilon$ is the *NESS effect* of $\mathcal{C} = c$ and the explanatory relation between the two is a NESS CAUSAL RELATION. The intuition behind NESS condition 1 is the same as the intuition behind but-for condition 1. The intuition behind NESS condition 2 is that for an *explanans* $\mathcal{C} = c$ to be *necessary* for the *sufficiency* of collection of *explanantia* to make $\mathcal{E} = \varepsilon$ true, there needs to be a collection of *explanantia* that are sufficient for $\mathcal{E} = \varepsilon$ (that collection being $\{\mathcal{C} = c, \bar{W} = \bar{w}\}$[20]) which can not be sufficient without $\mathcal{C} = c$. In other words, $\mathcal{C} = c$ is necessary for the sufficiency of $\{\mathcal{C} = c, \bar{W} = \bar{w}\}$.

$\bar{W} = \bar{w}$ is usually called a *witness*. □

At this point, one could propose to UNIFY the BUT-FOR CAUSATION with the NESS CAUSATION by showing that the explication of the former is a particular case of the explication of the latter. That can be done by proving that whenever the but-for subsumptive test is satisfied, then the NESS subsumptive test is satisfied as well. Note that even if such a proof is possible, it does not entail that the two explications of FACTUAL CAUSATION are UNIFIED in *every* system of concepts. For instance, it can be the case that one rejects this UNIFICATION for reasons of metaphysics or for rejecting that there is such thing as NESS CAUSATION in the first place. In what follows, I prove that there is at least one case where the but-for subsumptive test succeeds while the NESS subsumptive test fails.[21,22]

INTENTIONAL RULE. Not every *explanans* that passes the but-for test passes the NESS test.

PROOF: Assume a state of the world where Patrick throws a rock $\mathcal{R}_1$ that hits another rock $\mathcal{R}_2$ and due to the hit, the latter moves and hits a bottle breaking it.

$$Patrick \longrightarrow \mathcal{R}_2 \longrightarrow \text{Bottle}$$

$$\begin{aligned} Bottle \; e{=}\; & \mathcal{R}_2 \wedge U_3 \\ \mathcal{R}_2 \; e{=}\; & Patrick \wedge U_2 \\ Patrick \; e{=}\; & U_1 \end{aligned}$$

Patrick throwing $\mathcal{R}_1$ is a *but-for* cause according to the but-for subsumptive test. Specifically, according to facts of the case: $(U_1, U_2, U_3, Patrick, \mathcal{R}_2, Bottle) = (\mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T})$, $\mathcal{C} := Patrick$, $\mathcal{E} := Bottle$, $c = \mathcal{T}$, and $\varepsilon = \mathcal{T}$. The two but-for conditions hold:

- but-for condition 1: We want to test whether $\mathcal{M}, \bar{u} \models \mathcal{C} = c \wedge \mathcal{E} = \varepsilon$. I.e., whether $\mathcal{M}, (\mathcal{T}, \mathcal{T}) \models Patrick = \mathcal{T} \wedge Bottle = \mathcal{T}$. And it holds.;

tests fails to identify $\mathcal{C}$ as a NESS cause. This is why Beckers 2021 calls it "*direct NESS*". I do not provide Beckers's subsumptive test that identifies both *direct* & *in*direct NESS causes since it would have reduced the *EXPERT*-ORIENTED UNDERSTANDABILITY of my text without contributing anything more to my arguments even for the readers who have the expertise to understand its technicalities.

[20] My apologies for the abuse of notation. I do so to put emphasis on the intuition behind the formalism.

[21] I would like to thank dr. Sander Beckers for this counterexample as well as for their valuable feedback on my interpretation of their work both during and after my defence!

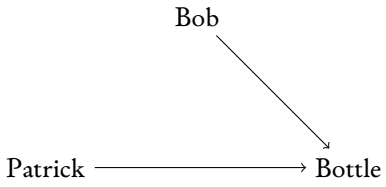
[22] It should be noted that Beckers's 2021 NESS subsumptive test that accounts for *both* direct & indirect causes (*see fn.* 19 above) *does* subsume BUT-FOR CAUSATION.

- but-for condition 2: We want to test whether $\exists c' \in \mathcal{R}(\mathcal{C})$ s.t. ($c' \neq c$ and $\mathcal{M}, \bar{u} \models [\mathcal{C} \leftarrow c']\neg\mathcal{E} = \varepsilon$). Indeed, for $c' = \mathcal{F} \in \mathcal{R}(Patrick) = \{\mathcal{T}, \mathcal{F}\}$, we have that $\mathcal{M}^{Patrick \leftarrow c'}, (\mathcal{T}, \mathcal{T}) \models \neg Bottle = \mathcal{T}$.

However, Patrick throwing $\mathcal{R}_1$ is *not* a *NESS* cause according to the NESS subsumptive test since the second NESS condition does not hold. The proof is left as an exercise to the reader! ;)

Now it is time to test whether NESS CAUSATION succeeds to identify causes in the case of CAUSAL OVERDETERMINATION.

EVALUATION: POSITIVE PARADIGMATIC EXAMPLE (OVERDETERMINATION).
• **Facts:** Both Patrick and Bob throw a rock and both rocks hit simultaneously the bottle breaking it in two different points.
• **Natural cause:** Both Patrick and Bob throwing a rock are natural causes of the bottle breaking.



$$Bottle \; e{=} \; (Patrick \vee Bob) \wedge U_3$$

$$Bob \; e{=} \; U_2$$

$$Patrick \; e{=} \; U_1$$

According to facts of the case: $(U_1, U_2, U_3, Patrick, Bob, Bottle) = (\mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T}, \mathcal{T})$. We want to test whether both $Bob = \mathcal{T}$ and $Patrick = \mathcal{T}$ are NESS causes of $Bottle = \mathcal{T}$. Due to symmetry, without loss of generalisation, I will only test whether $Patrick = \mathcal{T}$ is a but-for cause. Consequently, for this specific test, we have that $\mathcal{C} := Patrick$, $\mathcal{E} := Bottle$, $c = \mathcal{T}$, and $\varepsilon = \mathcal{T}$. Do the three NESS conditions hold?

- NESS condition 1: We want to test whether $\mathcal{M}, \bar{u} \models \mathcal{C} = c \wedge \mathcal{E} = \varepsilon$. Indeed, $\mathcal{M}, (\mathcal{T}, \mathcal{T}, \mathcal{T}) \models Patrick = \mathcal{T} \wedge Bottle = \mathcal{T}$ is true;
- NESS condition 2: We want to test whether there exists a $\bar{W} = \bar{w}$ s.t.
  - 2.α for every value $\bar{z}$ that the rest of the endogenous variables can take we have that $\mathcal{M}, \bar{u} \models [\mathcal{C} \leftarrow c, \bar{Z} \leftarrow \bar{z}]\mathcal{E} = \varepsilon$. There are two options for $\bar{W}$: $W = Bob$ or $W = \emptyset$
    - Let $W = \emptyset$: This entails that $Z = Bob$. For both possible values of z (i.e., $z = \mathcal{T}$ or $z = \mathcal{F}$), we have that $\mathcal{M}, \bar{u} \models [(Patrick, Bob) \leftarrow (\mathcal{T}, z)]Bottle = \mathcal{T}$ is true.
  - 2.β there exists a value $\bar{z}$ for the rest of the endogenous variables $\mathcal{Z} = \mathcal{V} \setminus \{W\}$ s.t. $\mathcal{M}, \bar{u} \models [W \leftarrow w, \bar{Z} \leftarrow \bar{z}]\neg\mathcal{E} = \varepsilon$.
    - For the case $W = \emptyset$: We have that $\bar{Z} = (Bob, Patrick)$. There exists a $\bar{z}$ ($\bar{z} = (\mathcal{F}, \mathcal{F})$) for which we have that $\mathcal{M}^{\bar{Z} \leftarrow \bar{z}}, \bar{u} \models \neg Bottle = \mathcal{T}$. □

  The NESS test is successful! Ergo, the explanatory relations are NESS CAUSAL RELATIONS. All of the following hold: $Bottle \; e{=} \; Partick$, $Bottle \; c{=} \; Partick$, $Bottle \; e{=} \; Bob$, and $Bottle \; c{=} \; Bob$!

We have verified that Beckers's 2021 formal explication of NESS CAUSATION satisfies the CORE PURPOSE of FACTUAL CAUSAL ATTRIBUTION in cases of CAUSAL OVERDETERMINATION! The process of formalisation does not stop here. More tests on PARADIGMATIC examples are needed to inquire whether we need to make further modifications. Finally, after exhausting the simplified PARADIGMATIC examples from formal philosophy of explanation, we need to move to the complex real-world examples from legal science (§III.3.3).

What's left to show is how the foregoing subsumptive tests can bee used to produce *causal justification* in legal ALGOAI.

IV.2.2 Subsumptive tests as CAUSAL JUSTIFICATIONS

In what follows, I sketch a methodology of employing the foregoing subsumptive tests to engineer legal ALGOAI that produces causal justifications for its output.

The first step is to parse through ECtHR's case-law, identify potential explanatory relations, and express them as structural equations. Then, we should test whether those relations satisfy the substantive tests of CAUSAL JUSTIFICATION (i.e., the subsumptive tests of BUT-FOR & NESS CAUSATION). Afterwards, we should express those equations using symbolic languages and can place *constraints* on connectionist AI like those described in §II.4.2.3. That can be done for instance by using logic programming languages of IF-THEN rules, where the IF-clause contains causes and the THEN-clause contains the effects (*cf.* Beckers 2021, p.6213; Wright 2011, p.289). Another way to do so is to represent the causal relations *via* graphs like those of Figure 1. They are *directed* graphs where the IF-clause of the structural equation is placed at the beginning of the arrow and the THEN-clause at its head (*cf.* Pearl 2022; Peters, Janzing, and Schölkopf 2017; Spirtes, Glymour, and Scheines 2000). We can represent such graphs by using the predicates `edge/2` & `node/1` in a logic programming language.[23] Note that so far, the causal relations represented by structural equations and causal graphs are *factual* causal relations. The ALGOAI models should also include structural equations and directed graphs that represent *legal* causal relations which as we saw in §1 are a *subset* of the former. Ergo, we need to identify more subsumptive tests to determine when a structural equation/an edge of a directed graph represents a legal causal relation and demarcate it as such (e.g., by introducing one predicate `edge_fact/2` to represent factual causal relations and one predicate `edge_legal/2` to represent a legal causal relation).

There is a final step missing though. We saw that a legitimate justification of legal ALGOAI should always connect the *laws* to the particular *facts* of each case (§I.3.2.1.2, ¶10). How does this happen in the case of structural equations? Well, since structural equations represent *laws* and since we can in general translate those laws in IF-THEN forms, we can incorporate them in the subsumptive-deductive justification paradigm we saw in §II.4.1.2. Take the example of the structural equation in §2, ¶6: $q\ e{=}\ p$, where $p :=$ "*The applicant denied the genocide.*" and $q :=$ "*The dignity of the victims of the genocide was harmed.*". The structural equation $q\ e{=}\ p$ can be construed as a general imperative of propositional functions $q(x)\ c{=}\ p(x)$ such that whenever for a particular a we have that $p(a)$ is the case, we get its effect $q(a)$. Note though that the two operators " $e{=}$" (causal structural equation) & "$\Leftarrow$" (deductive conditional) do not have the same rules of inference since one of them is a model of *causal* inference and the other one of *deductive* inference (*cf.* Woodward and Ross 2021, §2.5). More research on their differences is needed (*cf.* Iatrou 2022a, §4).

The proposed model of causal judicial justification is still not *legitimate* enough! We saw that causal laws can be construed as general imperatives extracted from the ECtHR's case-law. As one can see in the ECtHR judgements in HUDOC (ECtHR's database of past & ongoing cases), whenever the Court uses imperatives from its case-law, it always refers to the *sources* of that case-law. E.g., which past cases used the same imperatives? In which past cases one can find counterarguments and why those counterarguments are rejected? And so forth. This is a paradigmatic example of case-based reaosning (§I, *fn.* 24; §II.4.2.1). A way to incorporate the case-law sources from which we extrapolated the structural equations in causal models is introduced by Cabalar, Fandinno, and Fink 2014 (for its implementations in an actual logical programming language *see* Cabalar, Fandinno, and Muñiz 2020): we assign labels to causal relations that describe their content in natural language[24] and whenever a causal relation *fires* outputting its effect, the user can read in natural language that label and its contribution to the process of the input. Funnily enough, Cabalar, Fandinno, and Fink's proposal was motivated by designing logic-based XAI models of causal justifications so as to *attribute responsibility* in *law*.

IV.3 Conclusion: explicating explication

To conclude the last chapter of this Thesis, I would like to stress out once more the *non*-mechanistic *creative* nature of explication (§III.3.2, ¶2). As argued at the end of §III.3.2, most methods of engineering an explication should be be fleshed out during the *actual* practice of explicating and they should be tailor-made for the particularities of the explication cases at hand. Indeed, if one has closer look at the process of explicating in this chapter, they can abstract many repeating patterns that could potentially constitute methods of explicating forms of judicial justifications used for ALGOAI engineering (*cf.* §II.3.1.2.1, ¶4). For instance, a way to GLUE concepts is to identify which concepts realise the same CORE PURPOSES rendering them EXTENSIONALLY EQUIVALENT, or at least EXTENSIONALLY SIMILAR. This is also another step towards merging reflective equilibrium (RE) with explication (*cf.* §III.3.3). Moreover, a method of identifying the FORMAL FORM of a concept is to find the FOR-

[23] Literature on causal graphs in logic-based AI: Cabalar, Fandinno, and Fink 2014; Hyttinen, Eberhardt, and Järvisalo 2014; Zhalama et al. 2019; Peters, Janzing, and Schölkopf 2017, §7.2.1; Triantafillou and Tsamardinos 2015; Francis Rhys Ward and Belardinelli 2022; Rueckschloss and Weitkämper 2022. Literature on causal graphs in connectionist AI: Sivaram 2022; Beckers 2022; Wein et al. 2021; O'Shaughnessy et al. 2020. Since causal graphs force a specific *reaosning* structure in AI models, all of those citations can be construed as forms of XAI. Literature on how causal graphs can be used in computer science (or computational sciences) in general: Shafer 2002; Leslie 2002; Chen 2021; Kusner et al. 2017; Heinze-Deml, Maathuis, and Meinshausen 2018.

[24] The use of *natural language* is what makes the model of CAUSAL JUSTIFICATION a *semi*-formal model (*cf.* §III.1.1).

MAL FORMS of EXTENSIONALLY EQUIVALENT concepts and/or its HYPONYMS/HYPERNYNMS and perform the appropriate adjustments. In the foregoing examples, when the EXTENSIONALLY EQUIVALENT concepts are from different disciplines, what we have is cases of *comparative* METADI information being used to produce *contactual* METADI information (§II.3.1.2).

Concluding, the process of explicating judicial justifications concerns mainly the *designing* phase of engineering an ALGOAI model. It is about identifying *ex ante* all those components that we want a *legitimate* ALGOAI to have. It is then up to the AI engineers to make sure that the model will accommodate sufficiently those explications during the *building* phase. The last stage of engineering is that of *evaluating* the engineered ALGOAI model. Once more, legal scientists, logicians, & formal philosophers should make sure that the judicial justifications produced by the model the AI engineers built achieve the optimal COHERENCE with the PARADIGMATIC EXAMPLES of their disciplines. Only after it has successfully passed those coherence tests can the ALGOAI put into practice.

References

Alchourrón, Carlos E. 2015. "Limits of logic and legal reasoning." In *Essays in legal philosophy*, Reprint, edited by Carlos Bernal and Carla Huerta. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198729365.003.0017.

Aldrich, John. 1989. "Autonomy." *Oxford Economic Papers* 41 (1): 15–34.

Armstrong, D. M. 1983. *What is a law of nature?* Cambridge University Press.

———. 1997. *A world of states of affairs.* Cambridge University Press.

Askeland, Bjarte. 2015. "Basic questions of tort law from a Norwegian perspective." Chap. 2 in *Basic questions of tort law from a comparative perspective*, edited by Helmut Koziol, 99–164. Civil law. Jan Sramek Verlag. https://doi.org/10.26530/oapen_574832.

Baldwin, Richard. 2003. "A structural model interpretation of Wright's NESS test." Master's thesis, University of Saskatchewan.

Beckers, Sander. 2021. "The counterfactual NESS definition of causation." In *Proceedings of the 35th AAAI conference on Artificial Intelligence*, 35:6210–6217. 7. AAAI Press. https://doi.org/10.1609/aaai.v35i7.16772.

———. 2022. "Causal explanations and XAI." *Proceedings of Machine Learning Research 1st Conference on Causal Learning and Reasoning* 140:1–20.

Beckers, Sander, and Joost Vennekens. 2017. "The transitivity and asymmetry of actual causation." *Ergo* 4 (1): 1–27.

———. 2018. "A principled approach to defining actual causation." *Synthese* 195 (2): 835–862. https://doi.org/10.1007/s11229-016-1247-1.

Beebee, Helen. 2015. "Causes and laws: Philosophical aspects." In *International encyclopedia of the social and behavioral sciences*, 2nd ed., edited by James D. Wright, 266–273. Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.63007-6.

Black, Henry Campbell. 1968. *Black's law dictionary: definitions of the terms and phrases of American and English jurisprudence, ancient and modern.* Revised 4th ed. Edited by the publisher's editorial staff. St. Paul, Minn. West Publishing Co.

Braddon-Mitchell, David. 2017. "The glue of the universe." Chap. 6 in *Making a difference: Essays on the philosophy of causation*, edited by Helen Beebee, Christopher Hitchcock, and Huw Price, 99–115. Oxford University Press. https://doi.org/10.1093/oso/9780198746911.003.0006.

Brun, Georg. 2016. "Explication as a method of conceptual re-engineering." *Erkenntnis* 81 (6): 1211–1241. https://doi.org/10.1007/s10670-015-9791-5.

Cabalar, Pedro, Jorge Fandinno, and Michael Fink. 2014. "Causal graph justifications of logic programs." *Theory and Practice of Logic Programming* 14 (4-5): 603–618. https://doi.org/10.1017/S1471068414000234.

Cabalar, Pedro, Jorge Fandinno, and Brais Muñiz. 2020. "A system for explainable Answer Set Programming." *Electronic Proceedings in Theoretical Computer Science* 325:124–136. https://doi.org/10.4204/eptcs.325.19.

Canavotto, Ilaria. 2020. "Where responsibility takes you: Logics of agency, counterfactuals and norms." PhD diss., Institute of Logic, Language and Computation, Universiteit van Amsterdam.

Cartwright, Nancy. 1983. *How the laws of physics lie.* Oxford University Press.

———. 2002. "Against modularity, the causal Markov condition, and any link between the two: comments on Hausman and Woodward." *British Journal for the Philosophy of Science* 53 (3): 411–453. https://doi.org/10.1093/bjps/53.3.411.

———. 2003. "Two theorems on invariance and causality." *Philosophy of Science* 70:203–224.

———. 2007. *Hunting causes and using them: Approaches in philosophy and economics.* Cambridge University Press.

Cartwright, Nancy, and Sophia Efstathiou. 2011. "Hunting causes and using them: Is there no bridge from here to there?" *International Studies in the Philosophy of Science* 25 (3): 223–241. https://doi.org/10.1080/02698595.2011.605245.

Chen, Rong. 2021. "Causal network inference for neural ensemble activity." *Neuroinformatics* 19 (3): 515–527. https://doi.org/10.1007/s12021-020-09505-4.

Crimmins, James E. 2021. "Jeremy Bentham." In *The Stanford Encyclopedia of Philosophy*, Winter 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Davidson, Donald. 1967. "Causal relations." *Journal of Philosophy* 64 (21): 691–703.

de Jong, Willem R., and Arianna Betti. 2010. "The classical model of science: a millennia-old model of scientific rationality." *Synthese* 174 (2): 185–203. https://doi.org/10.1007/s11229-008-9417-4.

De Pierris, Graciela, and Michael Friedman. 2018. "Kant and Hume on causality." In *The Stanford Encyclopedia of Philosophy*, Winter 2018, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

de Vreese, Leen. 2006. "Pluralism in the philosophy of causation: Desideratum or not?" *Philosophica* 77 (1): 5–13. https://doi.org/10.21825/philosophica.82195.

Diakite, Mori. 2016. "Interventionism, realism and invariance: The kind of metaphysics that matter." Master's thesis, University of Oslo.

Dubber, Markus D. 2015. *An introduction to the Model Penal Code.* 1st ed. (online). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190243043.001.0001.

ECtHR Registry. 2021. *Guide on Article 10 of the European Convention on Human Rights: Freedom of expression.* Updated. April. https://www.echr.coe.int/documents/guide_art_10_eng.pdf.

ECtHR's Registry. 2022a. *Guide on Article 6 of the European Convention on Human Rights: Right to a fair trial (criminal limb).* Updated. August. https://www.echr.coe.int/documents/guide_art_6_eng.pdf.

———. 2022b. *Guide on Article 6 of the European Convention on Human Rights: Right to respect for private and family life, home and correspondence.* Updated. August. https://www.echr.coe.int/documents/guide_art_6_criminal_eng.pdf.

Fine, Kit. 1975. "Vagueness, truth and logic." *Synthese* 30 (3/4): 265–300.

Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli. 2022. "A causal perspective on AI deception." In *Proceedings of the International Conference on Logic Programming 2022 (ICLP 2022)*, edited by Joaquín Arias, Roberta Calegari, Luke Dickens, Wolfgang Faber, Jorge Fandinno, Gopal Gupta, Markus Hecher, et al. Haifa, Israel, July.

Galles, David, and Judea Pearl. 1997. "Axioms of causal relevance." *Artificial Intelligence* 97 (1–2): 9–43.

Gordon, Thomas F. 1988. "The importance of nonmonotonicity for legal reasoning." In *Expert systems in law: Impacts on legal theory and computer law*, edited by Herbert Fiedler, Fritjof Haft, and Roland Traunmüller, 4:111–126. Neue methoden im Recht. Attempto Verlag Tübingen GmbH.

Green, Sarah. 2015. *Causation in negligence.* Hart Studies in Private Law. Hart Publishing.

Gruyter, De, ed. 2013. "Critical essays on "Causation and responsibility"," https://doi.org/10.1515/9783110302295.

Hall, Ned. 2007. "Structural equations and causation." *Philosophical Studies*, no. 1, 109–136. https://doi.org/10.1007/s11098-006-9057-9.

Halpern, Joseph Y. 2000. "Axiomatizing causal reasoning." *Journal of Artificial Intelligence Research* 12:317–337.

———. 2016. *Actual causality*. The MIT Press.

Halpern, Joseph Y., and Judea Pearl. 2005. "Causes and explanations: A structural-model approach. Part I: Causes." *The British Journal for the Philosophy of Science* 56 (4): 843–887. https://doi.org/10.1093/bjps/axi147.

Hamer, David. 2014. "'Factual causation' and 'scope of liability': What's the difference?" *The Modern Law Review* 77 (2): 155–188. https://doi.org/10.1111/1468-2230.12063.

Hart, H. L. A., and Tony Honoré. 1959. *Causation in the law*. 1st ed. Oxford University Press.

———. 1985. *Causation in the law*. 2nd ed. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198254744.001.0001.

Heathcote, Adrian, and D. M. Armstrong. 1991. "Causes and laws." *Noûs* 25 (1): 63–73.

Heinze-Deml, Christina, Marloes H. Maathuis, and Nicolai Meinshausen. 2018. "Causal structure learning." *Annual Review of Statistics and Its Application* 5 (1): 371–391. https://doi.org/10.1146/annurev-statistics-031017-100630.

Hilpinen, Risto, and Paul McNamara. 2021. "Deontic logic: A historical survey and introduction." In *Handbook of deontic logic and normative systems*, edited by Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, vol. 1, Part I: Background, 3–136. College Publications.

Huber, Franz. 2013. "Structural equations and beyond." *The Review of Symbolic Logic* 6 (4): 709–732. https://doi.org/10.1017/S175502031300018X.

Hume, David. (1739) 2022. *A treatise of human nature*. Project Gutenberg, November 24, 2022. https://www.gutenberg.org/files/4705/4705-h/4705-h.htm#link2H_4_0086.

———. (1748) 2021. *An enquiry concerning human understanding*. Edited by Sir L. A. Selby-Bigge. Project Gutenberg, January 30, 2021. https://www.gutenberg.org/cache/epub/9662/pg9662-images.html.

Hyttinen, Antti, Frederick Eberhardt, and Matti Järvisalo. 2014. "dynamics of judicial-based causal discovery: conflict resolution with Answer Set Programming." In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, 340–349. UAI'14. Quebec City, Quebec, Canada: AUAI Press. https://doi.org/10.5555/3020751.3020787.

Iatrou, Evan. 2022a. "A normative model of explanation for binary classification legal AI and its implementation on causal explanations of Answer Set Programming." In *Proceedings of the International Conference on Logic Programming 2022 (ICLP 2022)*, edited by Joaquín Arias, Roberta Calegari, Luke Dickens, Wolfgang Faber, Jorge Fandinno, Gopal Gupta, Markus Hecher, et al. Haifa, Israel, July.

———. 2022b. "Non-monotonic rule-based logical programming for modelling legal reasoning: the example of Answer Set Programming." In *Proceedings of the 13th Panhellenic Logic Symposium*, 2:135–144. Volos, Greece, July.

Kadish, Sanford H. 1978. "Codifiers of the criminal law: Wechsler's predecessors." *Columbia Law Review* 78 (5): 1098–1144. https://doi.org/10.2307/1121892.

Kistler, Max. 2013. "The interventionist account of causation and non-causal association laws." *Erkenntnis* 78 (1): 65–84. https://doi.org/10.1007/s10670-013-9437-4.

Koziol, Helmut. 2015. "Comparative conclusions." Chap. 8 in *Basic questions of tort law from a comparative perspective*, edited by Helmut Koziol, 685–838. Civil law. Jan Sramek Verlag. https://doi.org/10.26530/oapen_574832.

Kusner, Matt J., Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. "Counterfactual fairness." In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 4069–4079. Long Beach, California, USA.

Lavrysen, Laurens. 2018. "Causation and positive obligations under the European Convention on Human Rights: A reply to Vladislava Stoyanova." *Human Rights Law Review* 18 (4): 705–718. https://doi.org/10.1093/hrlr/ngy027.

Law, Jonathan, ed. 2022. *A dictionary of law.* 10th ed. Oxford quick reference. Oxford University Press.

Leslie, Melanie B. 2002. "Liability for increased risk of harm: A lawyer's response to Professor Shafer." In *The dynamics of judicial proof: Computation, logic, and common sense*, edited by Marilyn MacCrimmon and Peter Tillers, Part VIII: Causality, 479–494. Studies in fuzziness and soft computing. Physica-Verlag.

Lombrozo, Tania. 2010. "Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions." *Cognitive psychology* 61 (4): 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002.

Lukaszewicz, Witold. 1990. *Non-monotonic reasoning: Formalisation of common-sense reasoning.* Edited by John Campbell. Hollis Horwood Series in Artificial Intelligence. Ellis Horwood.

Mackie, John L. 1965. "Causes and conditions." *American Philosophical Quarterly* 2 (4): 245–264.

———. (1976) 1986. *The cement of the universe: A study of causation.* Reprint. Clarendon library of logic and philosophy. Oxford University Press.

Menyhárd, Attila. 2015. "Basic questions of tort law from a Hungarian perspective." Chap. 4 in *Basic questions of tort law from a comparative perspective*, edited by Helmut Koziol, 251–344. Civil law. Jan Sramek Verlag. https://doi.org/10.26530/oapen_574832.

Mill, John Stuart. (1843) 1882. *A system of logic: Ratiocinative and inductive.* 8th ed. Harper & Brothers. https://www.gutenberg.org/cache/epub/27942/pg27942-images.html.

Moore, Michael S. 2009. *Causation and responsibility: An essay in law, morals, and metaphysics.* Oxford university Press. https://doi.org/10.1093/acprof:oso/9780199256860.001.0001.

———. 2019. "Causation in the law." In *The Stanford Encyclopedia of Philosophy*, Winter 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Morris, William Edward, and Charlotte R. Brown. 2022. "David Hume." In *The Stanford Encyclopedia of Philosophy*, Summer 2022, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Nolan, Donal. 2013. "Negligence and human rights law: The case for separate development." *The Modern Law Review* 76 (2): 286–318. https://doi.org/10.1111/1468-2230.12013.

O'Shaughnessy, Matthew, Gregory Canal, Marissa Connor, Mark Davenport, and Christopher Rozell. 2020. "Generative causal explanations of black-box classifiers." In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 5453–5467. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc.

Pearl, Judea. 2000. *Causality: Models, reasoning, and inference.* 1st ed. Cambridge University Press.

———. 2009. *Causality: Models, reasoning, and inference.* 2nd ed. Cambridge University Press.

———. 2022. *Causality: Models, reasoning, and inference.* Reprint. 2nd ed. Cambridge University Press.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: Foundations and learning algorithms.* The MIT Press.

Plakokefalos, Ilias. 2015. "Causation in the law of state responsibility and the problem of overdetermination: In search of clarity." *The European Journal of International Law* 26 (2): 471–492. https://doi.org/10.1093/ejil/chv023.

Press Service of the ECtHR. 2020. *Factsheet on hate speech.* Accessed October 28, 2021. https://www.echr.coe.int/Pages/home.aspx?p=press/factsheets&c.

Rigoni, Adam W. 2014. "Legal rules, legal reasoning, and nonmonotonic logic." PhD diss., University of Michigan.

Robertson, David W. 2009. "Causation in the Restatement (third) of Torts: Three arguable mistakes." *Wake Forest Law Review* 44 (4): 1007–1028.

Rueckschloss, Kilian, and Felix Weitkämper. 2022. "Correct causal inference in probabilistic logic programming." In *Proceedings of the International Conference on Logic Programming 2022 (ICLP 2022)*, edited by Joaquín Arias, Roberta Calegari, Luke Dickens, Wolfgang Faber, Jorge Fandinno, Gopal Gupta, Markus Hecher, et al. Haifa, Israel, July.

Russo, Federica. 2009. *Causality and causal modelling in the social sciences: Measuing variations*. Edited by Daniel Courgeau and Robert Franck. Vol. 5. Methodos Series. Springer.

———. 2023. "Causal pluralism and public health." In *The Routledge Handbook of philosophy of public health*, 98–113. Routledge. https://dare.uva.nl/search?identifier=6775285a-b960-426d-964d-7b30b9b07848.

Russo, Federica, and Benoît Rihoux. 2023. "Qualitative Comparative Analysis (QCA): A pluralistic approach to causal inference." Chap. 12 in *Oxford Handbook of Philosophy of Political Science*, edited by Harold Kincaid and Jeroen van Bouwell, Part 2: Methods in political science, debates and reconciliations. Oxford University Press.

Sartor, Giovanni. 2012. "Defeasibility in legal reasoning." Chap. 6 in *The logic of legal requirements: Essays on defeasibility*, edited by Beltrán Jordi Ferrer and Ratti Giovanni Battista, Part I: General features of defeasibility in law and logic, 108–136. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199661640.003.0007.

Sartorio, Carolina. 2009. "Causation and ethics." Chap. 26 in *The Oxford handbook of causation*, edited by Helen Beebee, Christopher Hitchcock, and Peter Menzies, Part VI: Causation in philosophical theories. Oxford University Press.

Shafer, Glenn. 2002. "Causality and responsibility." In *The dynamics of judicial proof: Computation, logic, and common sense*, edited by Marilyn MacCrimmon and Peter Tillers, Part VIII: Causality, 457–478. Studies in fuzziness and soft computing. Physica-Verlag.

Sivaram, Venkat, Abhishek; Venkatasubramanian. 2022. "XAI-MEG : Combining symbolic AI and machine learning to generate first-principles models and causal explanations." *AIChE Journal* 68 (6). ISSN: 0001-1541. https://doi.org/10.1002/aic.17687. https://browzine.com/articles/520116101.

So, Florence. 2020. "Modeling causality in law (Modélisation de la causalité en droit)." Master's thesis, University de Montreal (Faculty of Law), July.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 1993. *Causation, prediction, and search*. 1st ed. Vol. 81. Lecture Notes in Statistics. Springer-Verlag.

———. 2000. *Causation, prediction, and search*. 2nd ed. Adaptive Computation and Machine Learning. MIT Press.

Stapleton, Jane. 2009. "Causation in the law." Chap. 37 in *The Oxford handbook of causation*, edited by Helen Beebee, Christopher Hitchcock, and Peter Menzies, Part VII: Causation in other disciplines, 744–770. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199279739.003.0038.

Stoyanova, Vladislava. 2018. "Causation between state omission and harm within the framework of positive obligations under the European Convention on Human Rights." *Human Rights Law Review* 18 (2): 309–346. https://doi.org/10.1093/hrlr/ngy004.

———. 2020. "Common law tort of negligence as a tool for deconstructing positive obligations under the European Convention on Human Rights." *The International Journal of Human Rights* 24 (5): 632–655. https://doi.org/10.1080/13642987.2019.1663342.

Strasser, Christian, and G. Aldo Antonelli. 2019. "Non-monotonic Logic." In *The Stanford Encyclopedia of Philosophy*, Summer 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Strevens, Michael. 2008. "Comments on Woodward, *Making things happen*." *Philosophy and Phenomenological Research* 77 (1): 171–192.

Sulyok, Katalin. 2017. "Managing uncertain causation in toxic exposure cases: Lessons for the European Court of Human Rights from U.S. toxic tort litigation." *Vermont Journal of Environmental Law* 18:519–569.

Summers, Andrew. 2018. "Common-sense causation in the Law." *Oxford Journal of Legal Studies* 38 (4): 793–821. https://doi.org/10.1093/ojls/gqy028.

Triantafillou, Sofia, and Ioannis Tsamardinos. 2015. "Dynamics of judicial-based causal discovery from multiple interventions over overlapping variable sets." *Journal of Machine Learning Research* 16 (66): 2147–2205. http://jmlr.org/papers/v16/triantafillou15a.html.

Turton, Gemma. 2020. "Causation and risk in negligence and human rights law." *The Cambridge Law Journal* 79 (1): 148–176. https://doi.org/10.1017/S0008197319000898.

UN ILC (United Nations International Law Commission). 2001. "Draft articles on responsibility of states for international wrongful acts with commentarie." In *Yearbook of International Law Commission*, vol. II, bk. 2.

Wan, Hui, Michael Kifer, and Benjamin Grosof. 2015. "Defeasibility in answer set programs with defaults and argumentation rules." *Semantic Web* 6:81–98. https://doi.org/10.3233/SW-140140.

Wein, S., W. M. Malloni, A. M. Tomé, S. M. Frank, G. I. Henze, S. Wüst, M. W. Greenlee, and E. W. Lang. 2021. "A graph neural network framework for causal inference in brain networks." *Scientific Reports* 11 (1). https://doi.org/10.1038/s41598-021-87411-8.

West's Encyclopedia of American Law. 2004. 2nd ed. Gale. Accessed January 17, 2023. https://legal-dictionary.thefreedictionary.com/Model+Acts.

Widerker, David. 1985. "Davidson on singular causal sentences." *Erkenntnis* 23 (3): 223–242. https://doi.org/10.1007/BF00168290.

Woodward, James. 2003. *Making things happen: A theory of causal explanation*. Oxford University Press.

———. 2007. "Causation with a human face." Chap. 4 in *Causation, physics, and the constitution of reality: Russell's republic revisited*, edited by Price Huw and Corry Richard, 66–105. Clarendon Press.

———. 2015. "Methodology, ontology, and interventionism." *Synthese* 192 (11): 3577–3599. https://doi.org/10.1007/s11229-014-0479-1.

Woodward, James, and Lauren Ross. 2021. "Scientific explanation." In *The Stanford Encyclopedia of Philosophy*, Summer 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Wright, Richard W. 1985. "Causation in tort law." *California Law Review* 73:1735–1828.

———. 1988. "Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts." *Iowa Law Review* 73:1001–1077.

———. 2011. "The NESS account of natural causation: A response to criticisms." Chap. 14 in *Perspectives on causation*, edited by Richard Goldberg, 285–322. Hart Publishing.

Zhalama, Zhalama, Jiji Zhang, Frederick Eberhardt, Wolfgang Mayer, and Mark Junjie Li. 2019. "ASP-based discovery of semi-markovian causal models under weaker assumptions." In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 1488–1494. IJCAI'19. Macao, China: AAAI Press. https://doi.org/10.5555/3367243.3367245.

The Epilogue

This Thesis can be read as a *threefold manifesto*. Firstly, it is a manifesto about how SOCIETY 5.0 should accommodate (legal) ALGOAI in its political order. Secondly, it is a manifesto about how *logicians* and especially *formal philosophers* should reconceptualise their practice in the face of the emerging algocratic transdisciplinarity. It is a step towards Carnap's & Leitgeb's vision of a philosophy that escapes its academic sterilised isolation, a philosophy that intertwines itself with other disciplines opening new doors (Leitgeb 2013; Leitgeb and Carus 2022, §1 and §Supplement D). As well as a step towards Enlightenment's vision of a philosophy that has a *political & social* role, with formal philosophers & logicians becoming irreplaceable political authorities of SOCIETIES 5.0 in virtue of their expertise. Thirdly, it is manifesto on the importance of the nascent philosophy of science evolutionary stage. An attempt to turn the spotlight on what Mäki 2016 and many others envisaged as a *philosophy of metadisciplinarity*.

I would like to end this Thesis with highlighting a few research question that were left open. The most intriguing one being the response to the Ismene's dilemma in case that AI achieves an epistemic *agent* status. What would differentiate the position that a political order's functional dimension should be grounded on human reason from blatant crude *speciesism*? What about the right to *self-governance* of the new epistemic agent? And why should our information-processing method have a privileged status over theirs? A second open question to which I have a particular interest in is which should be the methodologies & policies that will allow judicial authorities and ALGOAI engineers to *check-&-balance* each other. Especially when it comes to *logicians & formal philosophers* checking-and-balancing judicial authorities for the *logical forms* of their justifications. Another research topic that I found rather compelling is the parallelism between the causal relations in a *legal* order and those in a *natural* order. For instance, which are their differences metaphysically and how those differences influence responsibility attribution? Concluding, a sum of open questions that are of importance for (legal) ALGOAI engineering are: *how* we can engineer AI models that can test whether Chinese-room generated judicial justifications reflect the judicial reaosning used by the human judicial authorities? *Which* other types of conceptual (re-)engineering can be used to model judicial justifications? *Can* we engineer *meta*-methodologies that compare different types of conceptual (re-)engineering? And finally, *which* should be the methodologies that will guide *meta*-disciplinary research?

Thank you for your patience.

References

Leitgeb, Hannes. 2013. "Scientific philosophy, mathematical philosophy, and all that." *Metaphilosophy* 44 (3): 267–275. https://doi.org/https://doi.org/10.1111/meta.12029.

Leitgeb, Hannes, and André Carus. 2022. "Rudolf Carnap." In *The Stanford Encyclopedia of Philosophy*, Fall 2022, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Mäki, Uskali. 2016. "Philosophy of interdisciplinarity: What? Why? How?" *European Journal for Philosophy of Science* 6 (3): 327–342. https://doi.org/10.1007/s13194-016-0162-0.

APPENDIX

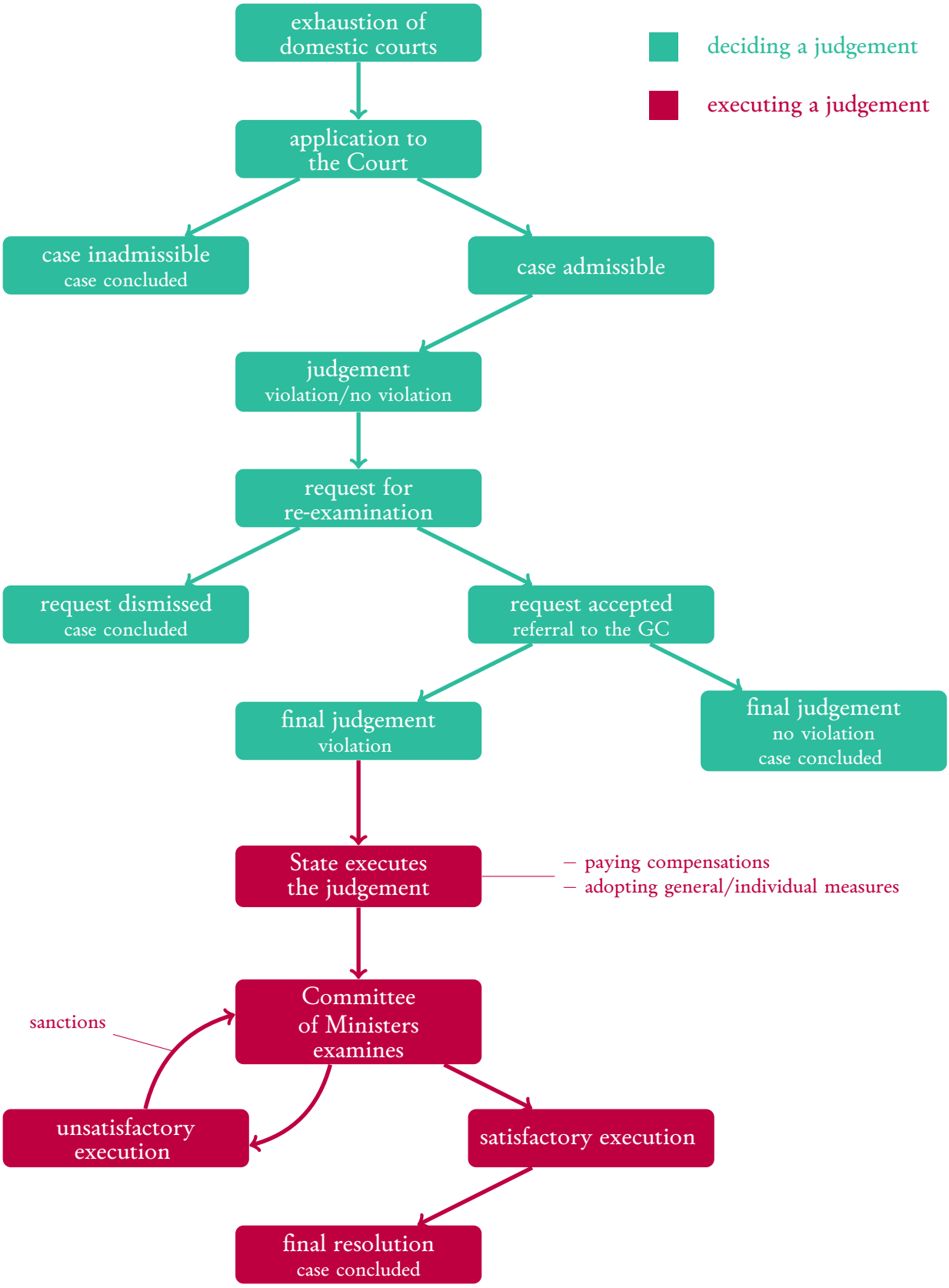


Figure 1: A visual representation of ***how the ECtHR's operates*** taken from the document "*The life of an application*" found in the ECtHR's website (accessed 23 February, 2023).

TABLE OF CASES

The ECtHR judges whether a state (aka *High Contracting Party* (*HCP*)) has violated the human rights of an individual (the *applicant*) that are protected by the European Convention of Human rights (the *Convention*). When I cite a case in the text, I will not cite its date unless it required by the context. Some cases have more than one dates (e.g., *Perincek v. Switzerland (2013) & Perincek v. Switzerland (2015)*) referring to different ECtHR judgements by different compositions of judges (e.g., Chamber & Grand Chamber respectively, where the latter overrides the judgements of the former). When I refer to cases with more than one dates without specifying the date, I am referring to characteristics of all of them. "GC" indicates that a judgement was decided by the Grand Chamber, the final ECtHR arbitrator. I will not include it unless it is required by the context.

..
Öneryıldız v. Turkey, Application no. 48939/99, 30 November 2004
Osman v United Kingdom Application No 23452/94, 28 October 1998

Perincek *v.* Switzerland, Application No. 27510/08, 17 December 2013
Perincek *v.* Switzerland [GC], Application No. 27510/08, 15 October 2015

Roman Zakharov v. Russia, Application no. 47143/06, 4 December 2015
Rostomashvili v. Georgia, Aapplication no. 13185/07, 08 November 2018
Rotaru v. Romania, Application no. 28341/95, 4 May 2000

Shvydika v. Ukraine, Application No. 17888/12, 30 October 2014
Soubeste and four other applications v. Austria and other States, Application no. 31925/22, Pending
Stomakhin *v.* Russia, Application No. 52273/07, 08 October 2018
Szabó and Vissy v. Hungary, Application no. 37138/14, 06 January 2016

Taganrog LRO and others v. Russia, pplications nos. 32401/10 and 19 others, 7 June 2022
Tyrer v. UK, Application No. 5856/72, 25 April 1978

Uricchiov v. Italy and 31 other States, application no. 14615/21, Pending

Volodina v. Russia (No.2), Application no. 40419/19,, 12 December 2019

Witzsch *v.* Germany, Application No. 7485/03, 13 December 2005

Convention articles

In what follows, I list the articles of the European Convention of Human Rights (the Convention) that I refer to in the Thesis. In come cases, I do not include paragraphs of articles that are not of relevance (e.g., ARTICLE 6's ¶¶2-3 in the French version). Note that whenever I refer to an article of the Convention in the Thesis, I am using SMALL CAPS to distinguish them from articles of other legal provisions.

ARTICLE 1
OBLIGATION TO RESPECT HUMAN RIGHTS

The High Contracting Parties shall secure to everyone within their jurisdiction the rights and freedoms defined in Section I of this Convention.

ARTICLE 2
RIGHT TO LIFE

1. Everyone's right to life shall be protected by law. [The rest of ARTICLE 2 is omitted.]

ARTICLE 3
PROHIBITION OF TORTURE

No one shall be subjected to torture or to inhuman or degrading treatment or punishment.

ARTICLE 6
RIGHT TO FAIR TRIAL

1. In the determination of his civil rights and obligations or of any criminal charge against him, everyone is entitled to a fair and public hearing within a reasonable time by an independent and impartial tribunal established by law. Judgment shall be pronounced publicly but the press and public may be excluded from all or part of the trial in the interests of morals, public order or national security in a democratic society, where the interests of juveniles or the protection of the private life of the parties so require, or to the extent strictly necessary in the opinion of the court in special circumstances where publicity would prejudice the interests of justice.

2. Everyone charged with a criminal offence shall be presumed innocent until proved guilty according to law.

3. Everyone charged with a criminal offence has the following minimum rights:

(a) to be informed promptly, in a language which he understands and in detail, of the nature and cause of the accusation against him;
(b) to have adequate time and facilities for the preparation of his defence;
(c) to defend himself in person or through legal assistance of his own choosing or, if he has not sufficient means to pay for legal assistance, to be given it free when the interests of justice so require;
(d) to examine or have examined witnesses against him and to obtain the attendance and examination of witnesses on his behalf under the same conditions as witnesses against him;
(e) to have the free assistance of an interpreter if he cannot understand or speak the language used in court

ARTICLE 6
DROIT À UN PROCÈS ÉQUITABLE

1. Toute personne a droit à ce que sa cause soit entendue équitablement, publiquement et dans un délai raisonnable, par un tribunal indépendant et impartial, établi par la loi, qui décidera, soit des contestations sur ses droits et obligations de caractère civil, soit du bien-fondé de toute accusation en matière pénale dirigée contre elle. Le jugement doit être rendu publiquement, mais l'accès de la salle d'audience peut être interdit à la presse et au public pendant la totalité ou une partie du procès dans l'intérêt de la moralité, de l'ordre public ou de la sécurité nationale dans une société démocratique, lorsque les intérêts des mineurs ou la protection de la vie privée des parties au procès l'exigent, ou dans la mesure jugée strictement nécessaire par le tribunal, lorsque dans des circonstances spéciales la publicité serait de nature à porter atteinte aux intérêts de la justice. [¶¶2-3 sont omis.]

ARTICLE 9
LIBERTÉ DE PENSÉE, DE CONSCIENCE ET DE RELIGION

[¶1 est omis.] 2. La liberté de manifester sa religion ou ses convictions ne peut faire l'objet d'autres restrictions que celles qui, prévues par la loi, constituent des mesures nécessaires, dans une société démocratique, à la sécurité publique, à la protection de l'ordre, de la santé ou de la morale publiques, ou à la protection des droits et libertés d'autrui.

ARTICLE 10
FREEDOM OF EXPRESSION

1. Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This Article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.

2. The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary

ARTICLE 14
PROHIBITION OF DISCRIMINATION

The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.

ARTICLE 17
PROHIBITION OF ABUSE OF RIGHTS

Nothing in this Convention may be interpreted as implying for any State, group or person any right to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms set forth herein or at their limitation to a greater extent than is provided for in the Convention

ARTICLE 22
ELECTION OF JUDGES

The judges shall be elected by the Parliamentary Assembly with respect to each High Contracting Party by a majority of votes cast from a list of three candidates nominated by the High Contracting Party.

ARTICLE 33
INTER-STATE CASES

Any High Contracting Party may refer to the Court any alleged breach of the provisions of the Convention and the Protocols thereto by another High Contracting Party.

ARTICLE 34
INDIVIDUAL APPLICATIONS

The Court may receive applications from any person, nongovernmental organisation or group of individuals claiming to be the victim of a violation by one of the High Contracting Parties of the rights set forth in the Convention or the Protocols thereto. The High Contracting Parties undertake not to hinder in any way the effective exercise of this right.

ARTICLE 41
JUST SATISFACTION

If the Court finds that there has been a violation of the Convention or the Protocols thereto, and if the internal law of the High Contracting Party concerned allows only partial reparation to be made, the Court shall, if necessary, afford just satisfaction to the injured party

OTHER LEGAL PROVISIONS

In what follows I provide a list of the legal provisions[1] cited in the Thesis. I exclude the references to the European Convention of Human Rights (ECtHR) since those can be found in other sections of the APPENDIX. Each entry begins with the abbreviation of the provision.

- CDL-AD(2011)003rev is the Venice Commission report titled "*On the rule of law*" that was adopted by the committee at its 86th session (Venice, 25-26 March 2011). Strasbourg, 4 April 2011, Study no. 512/2009.
- CDL-AD(2016)007 is the Venice Commission's "*Rule of law checklist*" adopted by the Venice Commission at its 106th plenary session (Venice, 11-12 March 2016). It is further endorsed by the Parliamentary Assembly of the Council of Europe on at its 4th session (11 October 2017), as well as by the Ministers' Deputies at the 1263rd meeting (6-7 September 2016) and the Congress of Local and Regional Authorities of the Council of Europe at its 31st session (19-21 October 2016). Strasbourg, 18 March 2016, Study No. 711/2013

- Resolution 1031 (1994). "*Honouring of commitments entered into by member states when joining the Council of Europe*". Adopted by the Parliamentary Assembly of the Council of Europe (PACE) on 14 April 1994 (14th Sitting). https://assembly.coe.int/nw/xml/XRef/Xref-XML2HTML-en.asp?fileid=16442

- UN A/HRC/38/35, "*Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*", UN's General Assembly, Human Rights Council 38th session, 18 June6 July 2018, Agenda item 3, Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development
- UN ILC 2001 (United Nations International Law Commission. 2001). "*Draft articles on responsibility of states for international wrongful acts with commentarie.*" In Yearbook of International Law Commission, vol. II, bk. 2

[1] I construe "*legal provision*" as an umbrella term that refers to any type of *authoritative* text (e.g., laws, treaties, international human rights instruments) whose authoritative content is about regulating the behaviour of a group of agents. That group of agents constitutes the *jurisdiction* of the legal provision. I base this construal on Governatori, Rotolo, and Sartor 2021, p.664.

The Enlightenment started with essentially philosophical insights spread by a new technology.

Our period is moving in the opposite direction.

It has generated a potentially dominating technology in search of a guiding philosophy.

How the Enlightenment ends
Henry A. Kissinger