LUND UNIVERSITY

AI Transparency in Trustworthy AI

From Metaphor to Governance Tool in EU Technology Regulation

Söderlund, Kasia

2025

Link to publication

Citation for published version (APA):
Söderlund, K. (2025). *AI Transparency in Trustworthy AI: From Metaphor to Governance Tool in EU Technology Regulation*. [Doctoral Thesis (compilation), Faculty of Engineering, LTH]. Department of Technology and Society, Lund University.

Total number of authors:
1

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

AI Transparency in Trustworthy AI

From Metaphor to Governance Tool in EU Technology Regulation

AI Transparency in Trustworthy AI

From Metaphor to Governance Tool in EU Technology Regulation

Kasia Söderlund

LUND
UNIVERSITY

DOCTORAL DISSERTATION
Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the Faculty of Engineering at Lund University to be publicly defended on 10 June 2025 at 9:15 in room E:1406, E-house, Department of Technology and Society, Klas Anshelms väg 12, Lund

Faculty opponent
Associate Professor Lena Enqvist

Organization: Lund University

Document name: Doctoral Dissertation

Date of issue 2025-06-10

Author: Kasia (Katarzyna) Söderlund

Sponsoring organization: Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS)

Title and subtitle: AI Transparency in Trustworthy AI: From Metaphor to Governance Tool in EU Technology Regulation

Abstract:

Transparency has emerged as a fundamental component of ethical AI guidelines around the world. In the European Union (EU), it is recognised as one of the core principles for fostering *Trustworthy AI,* and serves as a cornerstone in building an *ecosystem of trust* within the AI governance framework.

However, to support these ambitious policy objectives, the concept of transparency must be translated into clearly defined and implementable measures. Thus, by employing a combination of legal-doctrinal and socio-legal approaches, this compilation thesis aims to contribute to a clarified understanding of the concept of *AI transparency* in the EU's AI governance discourses. I examine the concept of AI transparency across four levels of abstraction: as a stand-alone objective, as a governance ideal, as a governance tool, and as a 'floating signifier'. Focusing in particular on AI transparency as a governance ideal and as a governance tool in relation to the EU's policymaking objective of Trustworthy AI, I analyse how AI transparency has been conceptualised, designed, and implemented for two stakeholder groups — individuals and oversight bodies — within the governance frameworks of the General Data Protection Regulation (GDPR), the Digital Services Act (DSA), and the Artificial Intelligence Act (AIA).

The main argument advanced in the thesis is that while AI transparency directed at individuals (understood as data subjects, service recipients, and natural persons) remains relevant, it is the effectiveness of oversight-oriented AI transparency that is crucial to the enforcement of the EU technology regulation and is, ultimately, foundational in the EU's pursuit of Trustworthy AI. Although transparency is central to the EU's vision for Trustworthy AI, its effectiveness depends on how legal obligations are interpreted, implemented, and enforced in practice.

Key words:

Transparency; Artificial intelligence; AI; AI Transparency; Trustworthy AI; European Union; EU; AI Act; Digital Services Act; General Data Protection Regulation

Supplementary bibliographical information

Language: English

Number of pages: 155

ISBN: 978-91-8104-549-9 (print), 978-91-8104-550-5 (electronic)

Recipient's notes

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature:

Date 2025-04-14

AI Transparency in Trustworthy AI

From Metaphor to Governance Tool
in EU Technology Regulation

Kasia Söderlund

LUND
UNIVERSITY

Coverphoto by Balise42, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Reflections_in_square_windows.jpg, edited by Kasia Söderlund in Affinity Photo 2.

Copyright pp 8-155 by Kasia Söderlund

Paper 1 © by Authors

Paper 2 © by Authors

Paper 3 © by Authors

Paper 4 © by Kasia Söderlund (Manuscript unpublished)

Faculty of Engineering

Department of Technology and Society

ISBN 978-91-8104-549-9 (print), 978-91-8104-550-5 (electronic)

Printed by E-husets Tryckeri, Lund University, Sweden

Lund 2025

To my family.

Table of contents

Abstract

Transparency has emerged as a fundamental component of ethical AI guidelines around the world. In the European Union (EU), it is recognised as one of the core principles for fostering *Trustworthy AI*, and serves as a cornerstone in building an *ecosystem of trust* within the AI governance framework.

However, to support these ambitious policy objectives, the concept of transparency must be translated into clearly defined and implementable measures. Thus, by employing a combination of legal-doctrinal and socio-legal approaches, this compilation thesis aims to contribute to a clarified understanding of the concept of *AI transparency* in the EU's AI governance discourses. I examine the concept of AI transparency across four levels of abstraction: as a stand-alone objective, as a governance ideal, as a governance tool, and as a 'floating signifier'. Focusing in particular on AI transparency as a governance ideal and as a governance tool in relation to the EU's policymaking objective of Trustworthy AI, I analyse how AI transparency has been conceptualised, designed, and implemented for two stakeholder groups — individuals and oversight bodies — within the governance frameworks of the General Data Protection Regulation (GDPR), the Digital Services Act (DSA), and the Artificial Intelligence Act (AIA).

The main argument advanced in the thesis is that while AI transparency directed at individuals (understood as data subjects, service recipients, and natural persons) remains relevant, it is the effectiveness of oversight-oriented AI transparency that is crucial to the enforcement of the EU technology regulation and is, ultimately, foundational in the EU's pursuit of Trustworthy AI. Although transparency is central to the EU's vision for Trustworthy AI, its effectiveness depends on how legal obligations are interpreted, implemented, and enforced in practice.

Popular science summary

It is not surprising that the public's perception of artificial intelligence (AI) is marked by ambiguity and confusion, as media coverage often portrays AI technologies through either utopian or dystopian narratives. In the utopian scenarios, AI is usually depicted as anything from friendly robots to a revolutionary technology capable of addressing today's most pressing challenges, including climate change, rising health costs, ageing population, optimising transportation, and developing more efficient public services. In contrast, the dystopian narratives present AI as a threat, often picturing it as killer robots or warning of AI superintelligence posing an existential risk to humanity — frequently illustrated with science fiction imagery, such as *The Terminator*.

Although much of AI's potential is exaggerated in sensational media news, and some consider it to be an overblown hype, AI is undoubtedly a powerful technology with remarkable capabilities. When well-trained, aligned with ethical values and legal standards, and applied in appropriate ways, AI can indeed be highly useful in many areas. However, it can also cause harm, as even a seemingly minor flaw in an AI system can have serious consequences for individuals, groups or even whole societies.

In the EU policymaking, the negative implications and risks posed by AI have been seen as a major obstacle in fully supporting innovation, development and uptake of AI technologies. At the core of the debate has been the need of both addressing the risks posed by the technologies, and supporting the ambition to harness AI's potential. The quest of realising both objectives at the same time has led the EU to develop the concept of *Trustworthy AI*. In essence, this policy approach means that AI deployed in the EU should be both *lawful*, *ethical* and *robust*. Within this framework, transparency is a key component enabling such objectives.

This thesis examines the concept of AI transparency in its various meanings within the AI governance discourses. While the exact meaning of this term may differ depending on the context, my focus is on exploring the way it operates on four levels of abstraction: as a stand-alone objective, as a

governance ideal, as a governance tool, and as – what I call – a 'floating signifier'. In particular, I compare how EU policymaking has envisioned AI governance as a governance ideal and how this vision has been articulated and embedded in EU regulations through various governance tools. To this end, I explore how AI transparency has been conceptualised, designed, and implemented in the General Data Protection Regulation (GDPR), the Digital Services Act (DSA), and the Artificial Intelligence Act (AIA) in relation to two stakeholder groups: individuals (EU citizens) and oversight bodies (national authorities or EU bodies responsible for enforcement of laws).

Under these EU legal frameworks, although individuals have been provided certain transparency rights, these are insufficient to achieve the objectives of Trustworthy AI on their own. However, national and EU oversight bodies have been granted a high level of transparency regarding AI systems on the basis of the above legal frameworks. This suggests that if any entity is in a position to ensure that AI systems deployed in the EU are lawful, ethical, and robust, it is the oversight bodies.

Thus, given that national authorities and EU bodies have now significant powers to monitor AI systems' compliance with all relevant laws, does this mean we have achieved the objective of Trustworthy AI? I argue that this is not necessarily the case, as the legal frameworks that have been introduced constitute only the first step in the pursuit of this goal. The way transparency provisions are implemented by organisations, utilised by oversight bodies, and the engagement of EU citizens in the AI governance framework will determine the actual effectiveness and strength of these rules. Among the most significant challenges is how public authorities exercise — or will exercise — their transparency mandates, which depends on various factors, such as their independence from political or industry pressures, their commitment to the enforcement responsibilities, and the resources at their disposal.

Although the oversight bodies have been entrusted the responsibility for enforcement of the EU regulations, the question remains how the Trustworthy AI framework will be implemented in practice. Looking beyond science fiction and utopian or dystopian narratives, it is important that it is the EU citizens who ultimately remain to be the governors in the Trustworthy AI governance framework.

Populärvetenskaplig sammanfattning

Det är inte förvånande att allmänhetens uppfattning om artificiell intelligens (AI) präglas av ambivalens och förvirring, eftersom mediernas rapportering ofta framställer AI genom antingen utopiska eller dystopiska berättelser. I de utopiska scenarierna framställs AI som allt från vänliga robotar till en revolutionerande teknik som kan lösa dagens mest akuta samhällsutmaningar, såsom klimatförändringar, stigande vårdkostnader, en åldrande befolkning, optimering av transporter samt utveckling av mer effektiva offentliga tjänster. I kontrast till detta presenteras AI i de dystopiska narrativen som ett hot – ofta i form av mördarmaskiner eller som en varning för superintelligent AI som utgör en existentiell risk för mänskligheten – ofta illustrerat med science fiction-bilder som från filmen 'The Terminator'.

Även om mycket av AI:s potential överdrivs i sensationssökande nyheter, och vissa betraktar tekniken som överdrivet upphaussad, är AI utan tvekan en kraftfull teknologi som kan öppna nya möjligheter inom många områden. När den tränas väl, är i linje med etiska värderingar och rättsliga normer, samt används i rätt sammanhang, kan AI vara mycket användbar. Men den kan också orsaka skada, eftersom även ett till synes mindre fel i ett AI-system kan få allvarliga konsekvenser för individer, grupper eller till och med hela samhällen.

De negativa implikationerna och riskerna som AI innebär för etiska värden och rättsliga ramar har inom EU:s policyutveckling betraktats som ett stort hinder för att fullt ut stödja innovation, utveckling och användning av AI-teknik. I centrum för debatten står behovet av att både hantera riskerna med teknologin och samtidigt främja ambitionen att tillvarata dess potential. Strävan efter att uppnå båda dessa mål har lett EU till att utveckla konceptet 'tillförlitlig AI'. I grunden innebär denna policyansats att AI som används inom EU ska vara laglig, etisk och robust. Inom denna ram är transparens en nyckelkomponent för att uppnå dessa mål.

I denna avhandling undersöker jag AI-systemens transparens inom EU:s policyvision för tillförlitlig AI. Även om begreppet 'AI-transparens' används

med olika betydelser inom AI-styrning, fokuserar jag i denna avhandling främst på två grundläggande sätt på vilka transparens tillämpas. Närmare bestämt jämför jag hur EU:s policyutveckling har föreställt sig AI-styrning som ett normativt ideal och hur denna vision har artikulerats och integrerats i EU:s reglering genom olika styrverktyg. I detta syfte undersöker jag hur AI-transparens har definierats, utformats och implementerats i den allmänna dataskyddsförordningen (GDPR), Digital Services Act (DSA) och AI-förordningen (AIA) i relation till två intressentgrupper: individer (EU-medborgare) och tillsynsorgan (nationella myndigheter eller EU-organ som ansvarar för rättslig tillsyn).

Under dessa EU-rättsliga ramar har individer visserligen tilldelats vissa transparensrättigheter, men dessa är inte tillräckliga för att på egen hand uppfylla målen för tillförlitlig AI. Däremot har nationella och EU-gemensamma tillsynsorgan beviljats en hög grad av transparens i fråga om AI-system, enligt de ovan nämnda rättsakterna. Detta tyder på att det i första hand är tillsynsorganen som har möjlighet att säkerställa att AI-system som används inom EU är lagliga, etiska och robusta.

Men givet att nationella myndigheter och EU-organ nu har betydande befogenheter att övervaka AI-systemens efterlevnad av relevanta lagar – innebär det att målet om tillförlitlig AI är uppnått? Jag hävdar att så inte nödvändigtvis är fallet, eftersom de rättsliga ramar som införts endast utgör ett första steg i denna strävan. Hur transparensbestämmelserna faktiskt implementeras av organisationer, hur de används av tillsynsorgan och hur EU-medborgarna engageras i styrningen av AI kommer att avgöra regelverkets faktiska genomslagskraft. En av de största utmaningarna är hur offentliga myndigheter utövar – eller kommer att utöva – sina transparensmandat, vilket i sin tur beror på faktorer som tillsynsaktivitetens omfattning och tillgång till resurser.

Även om tillsynsorganen har fått ansvaret att se till att EU:s regler efterlevs, kvarstår frågan om hur ramverket för tillförlitlig AI kommer att implementeras i praktiken. Bortom science fiction och utopiska eller dystopiska berättelser är det avgörande att det ytterst är EU-medborgarna som förblir styrande aktörer i governance-strukturen för tillförlitlig AI.

Streszczenie popularnonaukowe

Nie dziwi fakt, że postrzeganie sztucznej inteligencji (AI) przez opinię publiczną jest pełne niejasności i sprzeczności, ponieważ relacje medialne często przedstawiają technologie AI w narracjach albo utopijnych, albo dystopijnych. W scenariuszach utopijnych AI ukazywana jest zazwyczaj jako przyjazne roboty lub przełomowa technologia zdolna do rozwiązywania najpilniejszych problemów współczesnego świata, takich jak zmiany klimatyczne, rosnące koszty opieki zdrowotnej, starzejące się społeczeństwo, optymalizacja transportu czy usprawnienie usług publicznych. Z kolei w narracjach dystopijnych AI przedstawiana jest jako zagrożenie – często w formie zabójczych robotów lub jako superinteligencja stwarzająca egzystencjalne ryzyko dla ludzkości – często ilustrowane obrazami zaczerpniętymi z science fiction, jak np. z filmu *Terminator*.

Choć potencjał AI bywa w mediach przesadnie wyolbrzymiany, a niektórzy uważają go za przejaw nadmiernego „hype'u", nie ulega wątpliwości, że jest to technologia o ogromnym potencjale, otwierająca nowe możliwości w wielu dziedzinach. Odpowiednio wytrenowana, zgodna z wartościami etycznymi i normami prawnymi oraz stosowana we właściwym kontekście, AI może być bardzo użyteczna. Może jednak również wyrządzać szkody – nawet drobna wada w systemie AI może mieć poważne konsekwencje dla jednostek, grup społecznych, a nawet całych społeczeństw.

W unijnym procesie tworzenia polityk negatywne skutki i zagrożenia, jakie AI może stwarzać dla wartości etycznych i ram prawnych, zostały uznane za istotną przeszkodę w pełnym wspieraniu innowacji, rozwoju i wdrażania technologii AI. W centrum debaty znalazła się potrzeba jednoczesnego przeciwdziałania ryzykom wynikającym z technologii oraz wspierania ambicji wykorzystania jej potencjału. Dążenie do realizacji obu tych celów doprowadziło Unię Europejską do wypracowania koncepcji Godnej Zaufania AI (*Trustworthy AI*). W istocie ta polityka zakłada, że AI wdrażana w UE powinna być zgodna z prawem, etyczna i solidna. W tym podejściu przejrzystość stanowi kluczowy element umożliwiający osiągnięcie tych celów.

W niniejszej pracy analizuję rolę przejrzystości systemów AI w unijnej wizji politycznej Godnej Zaufania AI. Choć termin „przejrzystość" bywa rozumiany na różne sposoby w kontekście regulacji AI, moja praca skupia się przede

wszystkim na dwóch podstawowych aspektach, w jakich przejrzystość ta jest stosowana. Porównuję w niej, w jaki sposób unijny proces legislacyjny przedstawia zarządzanie AI jako ideał normatywny oraz jak ta wizja została ujęta i zakorzeniona w przepisach UE za pomocą różnych narzędzi regulacyjnych. W tym celu analizuję, jak pojęcie przejrzystości AI zostało skonceptualizowane, sformułowane i wdrożone w ogólnym rozporządzeniu o ochronie danych osobowych (RODO), Akcie o usługach cyfrowych (DSA) oraz Akcie o sztucznej inteligencji (AIA) w odniesieniu do dwóch grup adresatów: osób indywidualnych (obywateli UE) oraz organów nadzorczych (organów krajowych lub instytucji unijnych odpowiedzialnych za egzekwowanie prawa).

Na gruncie powyższych ram prawnych UE, choć osoby fizyczne uzyskały określone prawa do przejrzystości, to jednak same w sobie nie wystarczą one do osiągnięcia celów Godnej Zaufania AI. Z kolei krajowe i unijne organy nadzorcze otrzymały wysoki poziom dostępu do informacji o systemach AI, co wynika z wyżej wymienionych aktów prawnych. Sugeruje to, że jeśli jakieś instytucje są w stanie zapewnić, by systemy AI wdrażane w UE były zgodne z prawem, etyczne i solidne, to są to właśnie organy nadzorcze.

Czy zatem, skoro krajowe i unijne instytucje uzyskały znaczne uprawnienia do monitorowania zgodności systemów AI z obowiązującym prawem, można uznać, że cel wiarygodnej AI został osiągnięty? Twierdzę, że niekoniecznie, ponieważ wprowadzone ramy prawne stanowią dopiero pierwszy krok w realizacji tego celu. To, jak przepisy dotyczące przejrzystości będą wdrażane przez organizacje, wykorzystywane przez organy nadzorcze oraz jaką rolę odegrają obywatele UE w systemie zarządzania AI, przesądzi o faktycznej skuteczności tych przepisów. Jednym z największych wyzwań pozostaje to, w jaki sposób władze publiczne realizują – lub będą realizować – swoje mandaty w zakresie przejrzystości, co zależy od różnych czynników, takich jak poziom aktywności w działaniach nadzorczych, dostępne im zasoby, oraz niezależność od nacisków politycznych i branżowych.

Mimo że to organy nadzorcze ponoszą odpowiedzialność za egzekwowanie przepisów unijnych, nadal otwarte pozostaje pytanie, w jaki sposób ramy prawne dotyczące AI będą stosowane w praktyce. Odchodząc od narracji science fiction, utopii i dystopii, ważne jest, aby to właśnie obywatele UE pozostali ostatecznymi podmiotami sprawującymi władzę w systemie zarządzania Godną Zaufania AI.

List of Papers

Paper I

Explaining automated decision-making: a multinational study of the GDPR right to meaningful information.
Dexe, J., Franke, U., Söderlund, K., van Berkel, N., Jensen, R. H., Lepinkäinen, N., & Vaiste, J. (2022).
Geneva Papers on Risk and Insurance: Issues and Practice, *47*(3), 669–697.
https://doi.org/10.1057/s41288-022-00271-9

Paper II

Regulating high-reach AI: On transparency directions in the Digital Services Act.
Söderlund, K., Engström, E., Haresamudram, K., Larsson, S., & Strimling, P. (2024).
Internet Policy Review, *13*(1).
https://doi.org/10.14763/2024.1.1746

Paper III

Enforcement Design Patterns in EU Law: An Analysis of the AI Act.
Söderlund, K., & Larsson, S. (2024).
Digital Society 2024 3:2, *3*(2), 1–21.
https://doi.org/10.1007/S44206-024-00129-8

Paper IV

High-risk AI transparency? On qualified transparency mandates for oversight bodies under the EU AI Act.
Söderlund, K. (2025).
Technology and Regulation (accepted with minor changes).

Author's contribution to the Papers

Paper I

Jacob Dexe (JD), Ulrik Franke (UF), Kasia Söderlund (KS), Niels van Berkel (NvB), Rikke Hagensby Jensen (RHJ), Nea Lepinkäinen (NL), Juho Vaiste (JV)

JD and UF initiated, conceptualised the paper, and designed the research method. Each author was responsible for recruiting participants in their respective countries, and for gathering and analysing data. KS was responsible for conducting the study in Poland. JD and UF drafted the introduction, discussion and conclusions. KS, NvB, RHJ, NH, and JV contributed to the data analysis, findings, and discussion. KS identified a discrepancy in the interpretation of Article 15(1)(h) GDPR within the group, arising from differences in language versions of the regulation (discussed in the section 'Different interpretations and language versions of Article 15(1)(h) GDPR'). All authors reviewed the final manuscript.

Paper II

Kasia Söderlund (KS), Emma Engström (EE), Kashyap Haresamudram (KH), Stefan Larsson (SL), Pontus Strimling (PS)

KS was the primary author of the paper. EE and PS initiated the research idea, while SL, KH, and KS collaborated on the conceptual framework. EE authored Section 2 on recommender systems. KS wrote the introduction, the conceptual framework (in collaboration with KH), Section 3 on the legal analysis of the DSA, as well as the discussion and conclusions. All authors reviewed the final manuscript.

Paper III

Kasia Söderlund (KS), Stefan Larsson (SL)

SL and KS initiated the paper and jointly contributed to its conceptualisation and structure. KS, in collaboration with SL, authored the introduction,

conceptual framework, analysis, discussion, and conclusions. Both authors reviewed the final manuscript.

Paper IV

Kasia Söderlund is the sole author.

Other relevant publications

Söderlund, K. (2020). AI policy in Poland: Ethical considerations already at the core. In S. Larsson, C. I. Bogusz, & J. Andersson Schwarz (Eds.), *Human-Centred AI in the EU: Trustworthiness as a strategic priority in the European Member States* (pp. 66-84). European Liberal Forum asbl.

Larsson, S., Haresamudram, K., Högberg, C., Lao, Y., Nyström, A., & Söderlund, K. (2023). Chapter 39: Four facets of AI transparency. In *Handbook of Critical Studies of Artificial Intelligence*. Cheltenham, UK: Edward Elgar Publishing. https://doi.org/10.4337/9781803928562.00047

Larsson, S., Hildén, J., & Söderlund, K. (2025). Implications of Regulating a Moving Target: Between Fixity and Flexibility in the EU AI Act. (accepted for publication in *Law, Innovation and Technology*, 18.1)

Abbreviations

ADM	Automated decision-making
AI	Artificial Intelligence
AIA	Artificial Intelligence Act
CJEU	Court of Justice of the European Union
DPA	Data Protection Authority
DSA	Digital Services Act
DSC	Digital Services Coordinator
EDPS	European Data Protection Supervisor
EDPB	European Data Protection Board
EU	European Union
FOI	Freedom of information
GAI	Generative Artificial Intelligence
GDPR	General Data Protection Regulation
GPAI	General Purpose Artificial Intelligence
LLM	Large language model
ML	Machine learning
MSA	Market Surveillance Authorities
NB	Notified body
TEU	Treaty on the European Union
TFEU	Treaty on the Functioning of the European Union
VLOP	Very Large Online Platform
VLOSE	Very Large Online Search Engine

Acknowledgements

As I write the final words of this thesis, I'm filled with deep gratitude for the journey it represents. Being a PhD student is a rare privilege — the freedom of following one's interests and spending working days doing what one is most passionate about. At the same time, a PhD is not supposed to be easy or straightforward — at least it wasn't for me. But looking back now, I can see that every difficult moment, every frustrating struggle, was worth it. It all comes with the package and is part of the learning process. What kept me going, time and again, was the belief in the value and purpose of my work.

However, this PhD journey could have never been completed without the support of the many people I've worked with and learned from over the past five years. First of all, I've been fortunate to have Stefan Larsson as my main supervisor. I'm especially grateful that he introduced me to the world of socio-legal research, for his kind words during the most intense moments, and for always being there when I needed help. His support, patience, and positive approach made all the difference. I would also like to extend my sincere thanks to my second supervisor, Fredrik Heintz, for the discussions that challenged my own point of view, and which I very much hope to continue in the future.

I would also like to express my gratitude to two legal scholars who have inspired me to think beyond the 'legal box' and became academic role models for me. Thank you, Katja de Vries, for the presentation in Malmö back in 2018 that showed me how law and AI could be brought together. That talk led me to write my master's thesis about law and AI and, eventually, to embark on this PhD. I was therefore very happy and deeply grateful for her thoughtful guidance and encouragement at my midway review. I would also like to wholeheartedly thank Ida Koivisto, whose work has been the main inspiration for this thesis. It has been both an honour and a joy to have her as the reviewer at my final seminar, and her feedback gave me the courage to move forward with the idea for this thesis.

As a PhD student with legal background based at the Faculty of Engineering, I've truly appreciated the warmth and support from my colleagues at law faculties. I'm grateful to everyone I had the pleasure of meeting at the

Faculty of Law at Lund University during my regular visits. In particular, I wish to thank Vilhelm Persson for kindly inviting me for a study visit and for reviewing my draft manuscript. I also warmly thank Riikka Koulu for hosting me at the Faculty of Law at the University of Helsinki, and for the opportunity to meet the wonderful team at the Legal Tech Lab.

I am profoundly thankful to the Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS) for making this journey possible. The WASP-HS graduate school has been providing outstanding courses, inspiring summer and winter schools, conferences, and the study visit to Harvard University and MIT. I feel truly fortunate to have been part of such a unique research environment — one that brings together researchers from different backgrounds around the shared vision of AI serving our society. Many of these researchers have become my dear friends. A special thank you goes to our phenomenal programme director, Ericka Johnson, whose wholehearted dedication has made WASP-HS a truly supportive and meaningful research environment. Also, many thanks to Eva Sjöstrand, who has always been there for us — sometimes even outside working hours in cases of emergency.

I wish to warmly thank my home department, Technology and Society, and, in particular, Ingemar Bengtsson and Monika Baranowska, for their ongoing, kind support. My heartfelt thanks go also to my research group, AI & Society – Charlotte, James, Laetitia, Ellinor, and Kashyap. You've been the first to hear, challenge, and improve my ideas. Thank you for all the lively interdisciplinary conversations and being so supportive at all times.

But above all, I am deeply grateful to have such a wonderful family — both in Poland and in Sweden. Even worst days could be brightened up by my husband Jan – my foundation in life – and my precious twins, Helena and Alexander. They've been my greatest motivation for my work, and also my best distraction — thank you for helping me think about anything *but* AI once in a while. For months, you've patiently watched me work long hours, and asked when 'my book' would be ready. Now, at last, I can say: it's finished!

Lund, 23 April 2025

1 Introduction

Over the past three decades, significant advances in artificial intelligence (AI) research have led to a paradigm shift once deemed unattainable: humanity has created technologies that outperform its own cognitive and learning capabilities in many tasks[1]. AI technologies are now increasingly broadly adopted across both the private and public sectors, and new areas of application continue to emerge. In recent years, generative AI models — trained on vast datasets, often including personal data or copyrighted material — have become widely accessible to the general public (Bird et al., 2023; Hagendorff, 2024). In principle, where sufficiently large datasets exist, AI models can be developed and deployed.

Indeed, AI systems powered by machine learning (ML) algorithms process large datasets and learn autonomously, which enables their application across a wide range of settings. Generally, such systems are particularly well-suited in contexts involving handling large amounts of data, quick reaction times, or performing repetitive tasks (Dignum, 2019). Yet, the promise of AI is accompanied by significant concerns. One of the most pressing issues is the limited transparency about what exactly the systems learn during the training process, and the reasoning behind decisions made after their deployment. There are aspects of AI systems that are known to be uncertain, which can potentially be identified and tested. However, there are also 'unknown unknowns' — risks that cannot be anticipated in advance. This refers to situations, for example, when AI models behave unpredictably or fail in new contexts, on different datasets, or under real-world conditions, despite performing well on relevant testing datasets. This triggers a range of liability questions, since their reliability cannot be guaranteed by producers[2] to the same degree as with most conventional technologies. Moreover, the malfunctioning of the traditional technologies is usually immediately obvious, while AI harm can be subtle and may often

[1] The emergence of Artificial General Intelligence (AGI) — AI attaining human-level intelligence — is predicted on varying timelines by technology leaders (Browne, 2025; OECD, 2025).

[2] While the Product Liability Directive (PLD) (Directive 2024/2853) has been recently adopted, the AI Liability Directive (AILD) has been withdrawn by the Commission in February 2025 from its working agenda. See also the discussion in Section 6.5.

go unnoticed (Lu, 2024). Yet, when AI systems are deployed on scale, in particular in high-risk areas, even minor deficiencies in algorithmic operation can lead to disproportionate harm for individuals, social groups, or entire societies.

It could be argued that the technical challenges noted above can be addressed – at least to the extent possible – through more rigorous training, additional testing on relevant datasets, and by adopting appropriate risk management measures. However, beyond the technical complexities, the most significant reason for opacity in AI are often various types of legal barriers (Burrell, 2016; Pasquale, 2015; Tschider, 2021), particularly those based on the trade secrecy and business confidentiality rules. In many cases, algorithmic harms do not result from extremely complex AI, but from systems where transparency is deliberately limited.[3] As often pointed out in the scholarly literature, both technical and legal mechanisms are frequently intentionally employed to obscure algorithmic operations in order to evade external scrutiny — including from public authorities responsible for monitoring compliance with applicable laws (Burrell, 2016; Pasquale, 2015). The lack of transparency and accountability gaps have created conditions that allow algorithmic harms to proliferate (Lu, 2024), ranging from personal injuries to the amplification of societal inequalities.

Transparency challenges of AI systems are a well-established theme across policy documents, ethical frameworks, and in scholarly debates. In light of the concerns about the risks posed by AI, there is a widely shared recognition that transparency is a fundamental principle in ethical AI governance (see e.g. Jobin et al., 2019; OECD, 2024). However, despite this ostensible consensus, it is not a frictionless process when it comes to introducing specific, binding transparency measures. AI companies often resist disclosing how their systems are constructed, how they operate, and what objectives they are designed to achieve (Foss-Solbrekk & Glenster, 2022; Lu, 2024; Pasquale, 2010). Yet, in the absence of binding transparency requirements that enable external oversight, the AI industry is driven

[3] For example, in the Dutch automated decision-making system that led to the childcare benefit scandal in 2020, simple rule-based algorithms were used (van Bekkum & Borgesius, 2021; Wieringa, 2023). See also the case of an automated decision making system implemented in Gothenburg in 2020, resulting in incorrect school placement of hundreds of children (Kronblad et al., 2024).

primarily by market competition. This could incentivise companies to disregard legal and ethical obligations, resulting in a downward spiral of compliance and neglect of broader stakeholder concerns.

In the European Union (EU), AI technologies are seen as 'one the most strategic technologies of the 21st century' (European Commission, 2018a). However, the technical and legal barriers to AI transparency have been viewed as a major obstacle to addressing the issue of algorithmic harms and risks, and to adopting policy frameworks that fully support AI innovation. The European policymakers have therefore expressed strong commitment to ensuring that AI technologies used in the EU are compliant with all the binding laws, and are aligned with the EU values and principles. This twin objective – of addressing risks while supporting AI development – has resulted in the establishment of the overarching policy objective of *Trustworthy AI*, which means that only *lawful*, *ethical*, and *robust* AI systems may be deployed in the Union.

Within the Trustworthy AI framework, transparency features as one of its central components. Yet, as pointed to above, in spite of the seemingly universal agreement as to the importance of transparency in AI at the level of ethics guidelines and policymaking, delineating the exact scope of binding transparency rules remains to be a contentious issue.

Nevertheless, technology companies assert that the algorithmic systems they use are transparent to both end-users and regulators. For example, Meta's Transparency Center (Meta, n.d.) regularly releases numerous transparency reports on issues related to such areas as community standards, content restrictions, widely viewed content, along with the mandatory regulatory reports including *EU Digital Services Act: Systemic Risk Assessment Results Reports*.[4] Does it mean that Meta's services are transparent and trustworthy?

This thesis recognises that transparency in AI – or *AI transparency* – carry different meanings within AI governance discourses. In light of the tensions between both transparency and information secrecy, important questions

[4] As stated on Meta's website, 'As part of our ongoing commitment to transparency, we provide tools and information to help people understand Meta's technologies'– see Meta (2025). Meta's transparency reports are available at https://transparency.meta.com/reports/regulatory-transparency-reports/. Accessed 08/04/2025

are raised about who should have access to information about AI systems, when, and to what extent. Does 'AI transparency' mean that everyone should be able to see the source code of the system's model, have access to datasets and parameters? Should it be understood as an ethical principle or as legal obligations requiring disclosure of specific information? At what point can an AI system be considered transparent? This thesis explores the conceptual plurality of the term 'AI transparency' and seeks to clarify what meanings of AI transparency are at play in the AI governance discourses, particularly in relation to the EU's vision of Trustworthy AI.

1.1 Research aim

The overall aim of this compilation thesis is to contribute to a clarified understanding of the concept of *AI transparency* in the EU's AI governance discourses. The thesis conceptualises AI transparency across four levels of abstraction: 1) as a *stand-alone objective*, 2) as a *governance ideal*, 3) as a *governance tool*, and 4) as a *'floating signifier'*. The primary analytical focus under this aim is placed on examining AI transparency as a governance ideal and as a governance tool in relation to the EU's policymaking objective of *Trustworthy AI*.

The aim is studied within a selection of key EU technology regulations relevant for AI governance, namely the General Data Protection Regulation (GDPR), the Digital Services Act (DSA), and most recently adopted Artificial Intelligence Act (AIA).

The above aim is further operationalised in the following research questions:

RQ 1: How has individual-oriented AI transparency been conceptualised, designed, and implemented in the GDPR, DSA and AIA?

RQ 2: How has oversight-oriented AI transparency been conceptualised, designed, and implemented in the GDPR, DSA and AIA?

	Paper I	Paper II	Paper III	Paper IV
RQ1	x	x		
RQ2		x	x	x

1.2 Methodology and materials

In addressing the research questions above, this thesis adopts a combination of methodological approaches. At its core, the analysis is grounded in the *legal doctrinal method* (Watkins & Burton, 2018), which is central to both the appended papers and this thesis frame. The legal doctrinal method can be understood as 'research process used to identify, analyse and synthesise the content of the law' (Hutchinson, 2018). It is often seen as an intuitive aspect of legal work that 'forms the basis for most, if not all, legal research projects' (Watkins & Burton, 2018). However, since the concept of AI transparency is far from being contained within the realm of legal rules and provisions, the methodological lens that I apply is broader. Using the strict legal-doctrinal method alone for the investigation of the meaning of AI transparency would merely provide the answer to the question of how the concept has been expressed across legal frameworks.

As my interest has not only been to provide an account of how AI transparency has been designed in (a sample of) EU laws, the exploration of the various meanings of AI transparency was not limited to its legal meaning. Adoption of a broader conceptual context thus necessitates the application of research methods which go beyond the strictly doctrinal research. Indeed, one of the points of criticism to legal doctrinal approach is that the law itself acts as a theoretical structure that identifies and elevates certain specific facts as legally significant (Westerman, 2011). As a result, legal researchers often operate within a confined framework defined by law, not concerned with the effects of the law in the world external to the box labelled 'law' (Hutchinson, 2018). In other words, this approach may lead some researchers to believe that the law can be studied in isolation, without considering its social, political, moral, economic, or theoretical contexts. In response to these views, various new approaches to legal study

emerged, challenging the dominance of traditional doctrinal methods. One of such approaches which has become popular among legal academics over the last couple of decades is the socio-legal approach, also referred to as 'law in context' (Cownie & Bradney, 2017).

The approach applied in this thesis is therefore not limited to the legal doctrinal method, as it incorporates also *socio-legal* and *empirical* components. Although socio-legal studies are difficult to define due to the wide variety of research conducted under this label, they have been described in the literature as an 'approach to the study of law and legal processes which covers the theoretical and empirical analysis of law as a social phenomenon' (Cownie & Bradney, 2017). This approach therefore includes very diverse methods and perspectives, and also accommodates the range of methods adopted in this thesis.

Accordingly, Papers I-IV follow the legal doctrinal method, although they venture beyond the strictly doctrinal approach in different ways. The legal-doctrinal method could be seen as the point of departure in Papers II and IV, in which my position could be described as 'where do I see the expressions of AI transparency in these laws, and how are they phrased?'. In both articles, I identified and analysed the legal provisions relating to AI-governance. Paper II has examined the transparency rules relevant for the social media recommender systems used by very large online platforms (VLOPs) under the DSA, while in Paper IV, I followed the same steps with regard to the transparency mandates for oversight bodies in the AIA. In addition, as significant part of Paper II concerns the problems identified in governance of the social media recommender systems, the method could also be described as *contextual legal analysis* (Taekema & van der Burg, 2024).

In contrast, Papers I and III go beyond the legal doctrinal approach in different ways. The point of departure for Paper I was an *empirical method*. Empirical methods in law have been described as follows:

> Quantitative and qualitative empirical research into law and legal processes provides not just more information about law, (…) It answers questions about law that cannot be answered in any other way (Bradney, 2012).

In other words, interpreting legal rules alone may offer limited understanding as to how they actually operate in practice, and empirical

studies can help fill this gap (Burton, 2017). In most cases, however, in embarking on any empirical work concerning law, using doctrinal research to identify and interpret the relevant laws is necessary as well (Taekema & van der Burg, 2024). Thus, studying empirically the way transparency provisions are interpreted by the insurance sector in Paper I involved the legal analysis of the 'right to meaningful information' about decision-making processes in line with Art. 15(1)(f) GDPR[5].

In turn, Paper III involves the legal doctrinal approach as well in the analysis of the relevant provisions of the AI Act, yet the point of departure was *conceptual* instead. The article draws inspiration from the concept of *legal design patterns* (Koulu et al., 2021), focusing on how design thinking can inform the understanding and development of legal rules.[6]

Finally, the present thesis frame is based on the cross-cutting questions that connect the individual papers, and situates their findings within a broader socio-legal context. This approach enables the discussion on the conceptual development of transparency in AI governance, and provides a framework for a deeper reflection on the papers' findings.

The methodology adopted in this thesis could be described along three dimensions: legal-doctrinal, conceptual and temporal. The *legal-doctrinal dimension* focuses on the material scope of EU legal frameworks governing AI technologies. It investigates how the concept of transparency has been articulated within the selected EU regulations and examines the degree of information disclosure afforded to the stakeholders under the study.

Across the four papers, the choice of EU legal frameworks for this analysis is based on their significance in shaping AI governance in the EU:

- General Data Protection Regulation (GDPR) is a legal framework which is technology-neutral and is always applicable whenever personal data is processed. It is therefore relevant for AI systems which operate on personal data.
- Digital Services Act (DSA) has been applied in examining the legal governance of recommender systems (which is a type of AI

[5] The method is detailed in Section 4.1 and in Paper I.

[6] In brief, this approach could be described as identifying recurring structures in law which facilitates comparative analysis of the consequences of specific design choices in law.

technologies) used by social media platforms. This is an important aspect of AI governance, as recommender systems have significant impact through their personalisation features and their broad societal use.

- AI Act (AIA) is a technology-specific regulation that has been adopted to govern the use of AI systems in the EU. As Art. 1 AIA stipulates, its aim is to promote the 'uptake of human-centric and trustworthy artificial intelligence (AI)'. This legal framework is most recent out of the three; at the time of the defence of this thesis is in force, but is not fully applicable.

The second, *conceptual dimension* of this thesis engages with the conceptual entanglements of 'AI transparency'. It examines how this term operates across various levels of abstraction and interrogates the conceptual assumptions embedded in its use. The major source of inspiration in exploring the various meanings of 'AI transparency' was the insightful book *Transparency Paradox* by Ida Koivisto (2022). The author critically examines the concept of transparency, unpacking its use as a visual metaphor and challenging the common assumption that transparency inherently enhances institutional legitimacy. I have constructed a conceptual framework for understanding 'AI transparency' by adopting a similar approach. This framework explores the different meanings that may be at play behind the façade of the term 'AI transparency'. I understand AI transparency as a polymorphic and multi-dimensional concept that operates both as a stand-alone objective, a governance ideal, as a governance tool, and as a 'floating signifier' within the broader discourse on AI governance. While this account is by no means exhaustive or exclusive, it seeks to investigate some of the semantic work the term performs across contexts.

The *temporal dimension* is also important in this thesis as my PhD project has been carried out during a formative period for AI regulation in the EU. When I began this research project, the GDPR was the only binding legal instrument in place, and AI governance was still largely shaped by soft law and ethical guidelines. Since then, the regulatory landscape has evolved significantly. This thesis is therefore situated within the EU's shift from soft law approaches to binding legal frameworks. This transition is also reflected in the content of the papers: the conceptualisation, design, and, more

recently, implementation, of AI transparency is examined across the GDPR, DSA and AIA, with the progressing stages of the policy processes traced, or observed in real-time.

With regard to the materials used in this thesis — both in the included papers and this thesis frame — I draw on a range of materials to address the research questions and to situate the findings within a broader socio-legal context. The analysis engages with EU policy documents, adopted EU legal acts, and relevant academic and non-academic literature, including selected journalistic sources.

1.3 Delimitations

This thesis focuses on clarifying the various meanings of AI transparency within the EU's AI governance framework, yet its scope is subject to certain limitations. These limitations can be sorted along the legal-doctrinal, conceptual, and temporal dimensions as well.

With regard to the limitations in the *legal-doctrinal* dimension, the analysis primarily concerns how AI transparency is articulated in three EU legal frameworks: the GDPR, the DSA, and the AIA. The selection of these instruments has been explained earlier, as they address important aspects of AI governance in the EU. However, it is acknowledged that other EU and national legal instruments are also applicable to AI technologies, and some of them are referenced in this thesis. Among other legal frameworks that may be of significance for AI governance – depending on the context – are EU fundamental rights[7], data protection laws other than GDPR[8], consumer

[7] Charter of Fundamental Rights of the European Union (CFREU) protects rights of EU citizens, such as privacy, personal data, non-discrimination, and freedom of expression.

[8] For instance, for AI used in policing and criminal justice, the Law Enforcement Directive (LED) (Directive 2016/680) provides specific rules on data processing and profiling.

laws[9], competition laws[10], liability laws[11], intellectual property laws[12], cybersecurity laws[13], Data Governance Act[14], Data Act[15], and sectoral regulations, such as Medical Device Regulation (MDR) (Regulation 2017/745), which applies to AI used in medical devices (e.g. diagnostic software).

Although the focus in this thesis is on transparency concerning AI technologies, it also engages with the concept of automated decision-making (ADM). As further explained in Section 2.1.1, AI and ADM are overlapping but distinct concepts. In this thesis, the analysis of ADM is mainly referred to in relation to the GDPR' obligation to provide 'meaningful information', as explored in Paper I.

Moreover, the present study limits its analysis to AI transparency addressed towards two stakeholder groups: individuals (a term that I use herein as an umbrella term denoting data subjects, service recipients, natural persons, depending on the legal framework[16]) and oversight bodies (national authorities and EU oversight bodies responsible for enforcement of EU laws). AI transparency for other stakeholder groups, such as professional users of AI systems — e.g. hospitals, recruitment agencies, public

[9] For example, General Product Safety Regulation (GPSR) (Regulation (EU) 2023/988) concerns safety of non-food consumer products, including those using AI.

[10] For example, Digital Markets Act (DMA) (Regulation 2022/1925) targets large technology platforms ('gatekeepers'), limiting the use of AI for anti-competitive practices.

[11] The Product Liability Directive (Directive 2024/2853) governs liability for defective products, including AI systems.

[12] Trade Secrets Directive (Directive 2016/943) is often used by AI providers to protect AI software, training data, know-how, etc.

[13] Cybersecurity Act (Regulation 2019/881) addresses cybersecurity of digital products and services, introduces an EU-wide cybersecurity certification framework, potentially applicable to high-risk AI systems in future.

[14] Data Governance Act (Regulation 2022/868) regulates access to data, re-use of protected public-sector data, and concerns data intermediaries and data altruism mechanisms.

[15] Data Act (Regulation 2023/2854) sets out rules on who can access and use data generated by connected devices and services (such as smart devices, industrial machines, or cars). It focuses on non-personal data but also applies to mixed datasets (those containing both personal and non-personal data).

[16] In this thesis, I use the term *individual* to refer to an average person, without invoking the specific legal meanings associated with terms such as *data subject*, *service recipient*, *natural person*, or *EU citizen*.

authorities — are not in scope of this thesis, yet this could be subject of future research. The selection of individuals and oversight bodies is motivated by the fact that these groups occupy opposing ends of the AI transparency spectrum. Transparency for individuals reflects the most limited transparency level, while transparency for oversight bodies entails the most extensive form of transparency granted to third parties under the EU regulations.

The response to RQ 1. concerning the individual-oriented transparency is limited to the interpretation of Article 15 (1)(f) GDPR by the insurance sector as framed in Paper I, and to transparency obligations towards end-users, introduced by the DSA and studied in Paper II. Furthermore, oversight-oriented transparency in RQ 2. is analysed across Papers II – IV primarily with regard to the transparency mandates provided for oversight bodies under the DSA and the AIA. However, this thesis frame addresses the above limitations by analysing AI transparency from the perspective of both stakeholder groups across the three frameworks.

Although the questions concerning AI governance are relevant across jurisdictions around the world, the focus in this thesis is on the AI governance within the European Union. However, the scholarly literature, policy documents and journalistic sources I refer to are not limited to those of European authors. Moreover, it should be noted that the analysis of EU policy documents relating to the legal frameworks in scope of this thesis is based primarily on the EU Commission's documents.

With regard to the *conceptual limitations* – as already mentioned, this thesis does not suggest that the understanding of AI transparency can be constrained to the four meanings explored herein. However, the four conceptualisations have been developed for analytical purposes, drawing on Koivisto's (2022) critical examination of transparency as a concept. Other interpretations are conceivable and may complement or extend the approach taken in this study.

Concerning the *temporal limitations* – at the time Papers II–IV were written, both the DSA and the AIA were still in the early stages of adoption. Thus, the scope of their examination in the papers was limited to their textual analysis. In contrast, the GDPR was already applicable from 2018, which made it possible to investigate certain aspects of its implementation.

Nonetheless, this thesis frame also incorporates some of the enforcement updates concerning the GDPR and DSA, and briefly outlines the recent developments in the EU political landscape concerning AI governance.

1.4 Structure

This thesis is structured as follows. The present Chapter 1 provides a problem-oriented introduction, including the presentation of the research aim, applied methodology, materials, and delimitations.

Chapter 2 sets the stage by providing a broader contextual overview of AI technologies and how they have been approached by EU policymaking and laws. It first outlines relevant aspects of AI technological developments, their societal impact, and key issues which have attracted regulatory attention. It then turns to presenting the relevant EU policy documents and briefly introduces the legal frameworks examined in this thesis.

Chapter 3 engages with conceptual perspectives on transparency, presenting its meanings from an optical condition to a metaphor, and explains the metaphorical meanings of transparency as a governance ideal and a governance tool. Lastly, it is described how transparency has come to operate as a 'floating signifier'.

Chapter 4 provides an overview of the four papers that are included in this thesis. Each paper addresses an aspect of AI transparency in EU law. Paper I concerns the right of access to information for individuals under the GDPR. Paper II analyses transparency provisions oriented towards both individuals and oversight bodies under the DSA, in the context of social media platforms. Paper III analyses the enforcement mechanism in the AI Act, in light of the concept of *legal design patterns* and the enforcement framework of the GDPR. Paper IV investigates the role and limitations of oversight-oriented transparency under the AIA.

Chapter 5 presents the main analysis. It brings together the conceptual perspectives on transparency introduced in Chapter 3 and insights from the Papers to explore four different meanings of AI transparency that are at play in the EU's AI governance discourses. These are conceptualised in this thesis as 1) AI transparency as a *stand-alone objective*, 2) AI transparency as a

governance ideal, 3) AI transparency as a *governance tool*, and 4) AI transparency as a *'floating signifier'*. The analysis focuses on addressing the research questions by exploring how the concept of AI transparency – as a governance ideal and a governance tool – has been conceptualised, designed, and implemented across the GDPR, DSA, and AIA frameworks in relation to individuals and oversight bodies, respectively.

Chapter 6 moves into a discussion of the findings in light of the EU's Trustworthy AI objective. It briefly revisits the four meanings of AI transparency considered in this thesis, reflects on the limited scope of transparency for individuals, the reliance on oversight in ensuring effectiveness of the AI governance framework, and notes the observable political narrative shift concerning AI regulation in the EU in the recent months.

Chapter 7 concludes the thesis by summarising the main findings.

2 Context: AI technologies meet EU regulation

Artificial intelligence (AI) has evolved from a theoretical idea to one of the most revolutionary technologies of our times. However, with increasingly widespread AI applications come a range of risks and negative effects. In the European Union (EU) context, AI technologies have been seen both as a strategic opportunity and a source of potential harm — triggering regulatory measures which aim at balancing innovation with the protection of fundamental rights and public interest.

In this chapter, I set the scene for examining the role of transparency in EU's AI governance by providing an overview of key developments in AI field, and how the EU policymakers have responded to opportunities and risks of AI. The chapter concludes with a brief overview of the three legal frameworks – the GDPR, the DSA and the AIA – that have been adopted as part of the EU's policy approach to AI governance, and which also form the focus of this thesis.

2.1 The technology

The nomenclature of artificial intelligence opens doors, attracting investors, media attention, and excitement. Today, its popularity is so pervasive that 'the label of AI is being slapped onto nearly any piece of code' (Solove, 2024). However, despite the widespread use of these technologies, there is a lot of confusion about the capabilities and limitations of AI systems. Some critics argue that the term 'artificial intelligence' itself is misleading, as AI systems are neither *artificial*[17], nor meaningfully *intelligent*[18]. Other critics,

[17] Rather, as Crawford (2021) argues, AI is 'both embodied and material, made from natural resources, fuel, human labor, infrastructures, logistics, histories, and classifications.'

[18] Although AI systems can outperform humans in certain tasks, the algorithms merely *simulate* intelligence – they do not understand in any meaningful sense (Mitchell, 2019).

such as Floridi (2024), make an argument that the characteristics of the current hype surrounding AI reflect familiar patterns associated with technological bubbles.

However, AI systems and other algorithmic technologies are already deeply embedded in our societies and everyday lives. From recommender systems and chatbots to AI-assisted medical diagnostics and autonomous vehicles, their use is widespread and growing. While AI technologies may indeed be overhyped, their negative impacts are already manifest and should not be overlooked.

2.1.1 The emergence of AI and automated decisions

Although the scientific field of artificial intelligence (AI) has been around at least since the 1950s[19], the first truly viable forms of AI have emerged only during the past three decades (Cristianini, 2021). The substantial breakthroughs in machine learning (ML), the massive datasets and computing power had to coevolve before they could benefit from each other.

Much of the early AI research focused on explicitly encoding human knowledge as formal rules and facts, and on variations of logical reasoning, often using 'if-then' logic. This so-called *symbolic* or *top-down* approach was a dominant paradigm for decades (Dignum, 2019).[20] It led to impressive achievements in certain areas, yet such systems would often struggle with tasks involving implicit or common-sense knowledge, and failed to prove useful in tasks in which knowledge was not possible to be captured and formalised by rule-based approach (Dignum, 2019). [21]

[19] Most AI researchers trace the field's official beginning to a 1956 workshop at Dartmouth College, led by John McCarthy (Mitchell, 2019).

[20] The *symbolic* approach is also known as GOFAI (Good Old-Fashioned AI). This method has seen success in expert systems and rule-based programs like IBM's DeepBlue, which was the first machine-based system that defeated world chess champion Garry Kasparov in 1997. However, unlike modern AI, Deep Blue relied on brute-force computation and explicit programming to evaluate possible moves (Dignum, 2019).

[21] The field has experienced periods of intense innovation followed by setbacks – often referred to as 'AI winters' (Russell & Norvig, 2021) – where enthusiasm waned due to technological limitations.

Yet, a major paradigm shift occurred in the 1990s, as AI research moved away from rule-based systems toward data-driven *machine learning* (ML) (Russell & Norvig, 2021). These approaches – often called *sub-symbolic* or *bottom-up* methods (Dignum, 2019) – do not follow the rule-based logic. Instead, they learn from massive datasets and past experience to, for example, identify patterns in data and produce increasingly accurate predictions[22]. To achieve such capabilities, the algorithms discern relevant features of the data that usually are not obvious, intuitive, or even explainable to humans. This transition enabled AI to tackle increasingly complex tasks (see e.g. Dignum, 2019).

A significant subcategory of machine learning is *deep learning*, which deals with the development and application of deep neural networks. These systems, taking inspiration from how the human brain works, are generally seen as better suited for tasks involving uncertainty, perception, and pattern recognition. Such artificial neural networks are optimised and trained for specific tasks, and they can differ profoundly in terms of their architecture and mode of operation (Steimers & Schneider, 2022).

Current advancements in AI are primarily driven by the capabilities of machine learning, and deep learning methods in particular. As Melanie Mitchell (2020) observes, deep learning (or deep neural networks) methods have become the dominant AI paradigm to the extent that in much of the popular media the term 'AI' itself has come to mean deep learning.

However, as pointed out, these approaches rely heavily on the availability of vast volumes of data and computational resources.

Platformisation and datafication

After 2010s, the scale of personal data sharing and collecting has increased dramatically, due to what Poell et al. (2019) refers to as the process of *platformisation*. Globally operating digital platforms are becoming increasingly central to public and private life, utilising massive datasets for

[22] One of the first most iconic moments showcasing the capabilities of this approach is Google DeepMind's AlphaGo, which defeated Go world champion Lee Sedolin 2016. Instead of relying on pre-programmed strategies, AlphaGo trained itself by analysing vast amounts of game data, recognising patterns, and improving its performance over time (see, for instance, Gibney, 2016; Yeung & Ranchordás, 2024).

training of AI models. The platformisation as a trend is itself linked to the process of *datafication*, referring to the ways in which digital platforms render into data 'practices and processes that historically eluded quantification' (Poell et al., 2019). The process involves users' explicitly shared personal data, but also behavioural meta-data, collected through expanding platform infrastructures in form of apps, sensors, and trackers embedded in various personal devices. Moreover, the widespread societal use of generative AI has intensified data collection, as users increasingly share sensitive information during interactions with these systems (Sebastian, 2023).

Today, technology allows data collecting organisations to make use of personal data on an unprecedented scale in order to pursue their economic activities.

Defining AI

In light of the above developments in the AI field, it could be seen that the notion of 'AI' includes a broad set of approaches (Mitchell, 2019). It spans many domains, such as planning, problem-solving, communication, language comprehension and pattern recognition. Since each research area approaches their objectives differently, many AI sub-fields often have little in common (Dignum, 2019).

Given the wide range of methods and technologies associated with AI, it is unsurprising that there is no straightforward or universally agreed definition of AI. [23] In the AI field, discussions continue over how to define the scope of AI and what should or should not fall under that label (Lemley & Casey, 2019). Since this thesis does not attempt to resolve this debate, suffice is to say that AI is a broad and evolving concept. As Kate Crawford (2021) suggests, this very broadness of the term – its ambiguity, malleability and openness — is precisely what allows it to be adapted and applied in many ways, depending on the context.

[23] In very broad terms, AI is described as the development of machine-based systems that perform tasks typically requiring human intelligence (Dignum, 2019) or as 'a set of techniques aimed at approximating some aspect of human or animal cognition using machines '(Calo, 2017). In a somewhat more specific way, Russell & Norvig (2021) describe AI as 'machines that can compute how to act effectively and safely in a a wide variety of novels situations' (p.19).

However, for regulatory purposes, the plurality of techniques and ways of understanding of AI constitutes a problem, as the object of regulation needs to be clearly specified. This point is of key importance, as the definition of an 'AI system' determines which computational systems and processes are within the scope of legal obligations, limitations and rights. This is why the definition of AI during the legislative process on the AI Act was one of the most thorny issues (see, for example, Ruschemeier, 2023).

The definition of artificial intelligence in the AI Act, which will be analysed below, was eventually agreed upon[24] and included in the main text of the AI Act. In Recital 12, the AI Act states that 'the definition should be based on key characteristics of AI systems that distinguish it from simpler traditional software systems or programming approaches and should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations'. Accordingly, the definition stipulated under the AI Act in Article 3 (1) defines AI systems as follows:

> 'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

The definition is, admittedly, rather complex. To make sense of it, it helps to break it down into a few essential – cumulative or optional – components. In scope of the AIA are thus systems which:

1. process inputs (by using techniques which enable them to learn from the input, reason and/or model) to create certain outputs — such as predictions, recommendations, or decisions,
2. by operating towards explicit or implicit objectives,
3. that work with various levels of independence (i.e. making decisions with varying degree of involvement of humans, including fully autonomous operation).

And optionally:

[24] The AI Act definition of AI has been aligned with the definition laid down by UNESCO. As Recital 12 AIA state, the definition of an 'AI system' should be 'closely aligned with the work of international organisations working on AI to ensure legal certainty, facilitate international convergence and wide acceptance'.

4. *may* have self-learning capabilities after being deployed (allowing the systems to continue learning after deployment),
5. depending on their autonomy degree, they *can* influence physical or digital environments.

Many automated systems would not meet the cumulative points above. They would therefore be only categorised as automated (with only rule-based algorithms), autonomous systems (if there is no element of creating new outputs), or which do not have impact on the external environment (physical or virtual), and thus fall outside of scope of the AI Act[25] (but still be subject to other laws).

While the primary focus of this thesis is on AI, Paper I specifically examines automated decision-making (ADM) – a concept that may include but is not limited to AI. The concept of automated decision-making (ADM) will therefore be explained in the forthcoming section.

Automated decision-making (ADM)

As mentioned, AI is often discussed alongside automated decision-making (ADM), and the terms 'ADM' and 'AI' may be understood as overlapping concepts. Yet, as Lomborg et al. (2023) observe, there is often a conceptual confusion about what ADM actually includes — and where its boundaries lie.

In general, ADM covers a wide range of systems — from advanced AI to simple automated systems, when used in contexts that shape decisions and outcomes affecting individuals. For example, Algo:aware report published in 2018 – procured by the European Commission – defines the term as 'a software system (…) that autonomously or with human involvement, takes decisions or applies measures relating to social or physical systems on the basis of personal or non-personal data, with impacts either at the individual or collective level' (algo:aware, 2018). Also Richardson (2022) in her definition of ADM emphasises their impact on humans, describing ADMs as 'any systems (…) that use computation to aid or replace government decisions (…) that impact opportunities, access, liberties, rights, and/or

[25] It should also be noted that the scope of the definition is also subject to significant carveouts on the basis of Art. 6 AIA.

safety.' This definition includes fully automated systems as well as decision-support systems that combine automated aspects with the involvement of human decision-making.

Thus, the concept of ADM can be understood as the processes of implementing and delegating tasks to automated systems and includes both rule- and knowledge-based algorithms (Lomborg et al., 2023). Crucially, there is the human decision component in the equation, which influences the way how ADM systems are shaped, interpreted, and used (Lomborg et al., 2023).

2.1.2 AI opacity and risks

As the scepticism surrounding AI's potential has now largely faded, AI technologies are being increasingly deployed across both private and public domains. AI applications are now used in such sectors as healthcare (e.g. in identifying a variety of eye and skin disorders, detecting cancers (Högberg, 2025)), transportation (e.g. autonomous cars), in employment, education, public services, agriculture, in fraud detection (e.g. Dignum, 2019; Littman et al., 2021). The deep-learning features have enabled AI to be applied for such complex tasks as sound, voice and speech recognition, image processing and facial recognition, language translation, and content generation (Littman et al., 2021).

In view of such remarkable AI capabilities, it could be argued that if AI did not pose any concerns to issues such as safety, fundamental rights, liability, or competition, there would be no need for regulation.[26] However, regulation by law is often seen as a necessary tool in balancing the conflicting rights and interests (Lessig, 1999). In the context of AI technologies, this tension typically emerges between, on one hand, the freedom of AI providers to conduct business, and on the other, the rights and interests of various other stakeholders. On the latter side, among the most often recurring themes of risks and negative effects of AI are such issues as challenges to due process principles, liability, privacy and

[26] Notably, regulation by law is not the only method of shaping societal behaviour. As Lessig (1999) argues, human behaviour may be regulated by four modalities: through law, social norms, architecture and by market forces.

autonomy, mental health of users or in some cases even their physical safety. On the larger, societal scale, the issues often pointed to include unfair competition, the public security concerns, dissemination of harmful or illegal content, and the negative impact on political discourse (see e.g. Fortes et al., 2022; Galaz et al., 2021; Kaminski, 2022; Scherer, 2015).

Many of the above challenges are seen as stemming from a common, underlying problem: the opacity of AI systems. However, the notion of opacity in AI needs to be further unpacked.

In general, opacity of AI systems may arise from different reasons. Jenna Burrell (2016), for example, distinguishes opacity as an intentional corporate or institutional self-protection, opacity resulting from technical illiteracy and as opacity that stems from algorithmic complexity exceeding the cognitive abilities of humans. I will outline the various reasons of AI opacity below, yet I will divide these by grouping into technical opacity, legal opacity, and opacity due to the high reach and personalisation of certain AI systems.

Technical opacity

As mentioned, the lack of transparency of AI systems often stems from their inherent technical complexity (Burrell, 2016; de Laat, 2017; Pasquale, 2015). In particular, the internal workings of deep learning algorithms, or the interplay between multiple algorithms, makes the scrutiny of such systems a task often extending the human cognitive capacity. This issue is commonly referred to as the *black-box* problem (Pasquale, 2015).

Since the most successful AI systems applying deep learning are effectively 'black-boxes' for humans, the issue that is often pointed to in this regard is the uncertainty about what these systems actually learn. Supervised, unsupervised, or reinforcement learning methods are optimised for tasks such as classification or pattern recognition, and the models themselves function by identifying correlations – not causation, yet the patterns they detect may appear to suggest otherwise. Although rigorous testing may

help to mitigate this issue, many problematic AI operations can still go unnoticed[27].

In response to the 'black-box' problem in AI, research areas such as *eXplainable AI* (XAI) have evolved with the aim to develop more interpretable AI models and providing ways to decipher opaque AI algorithms. Moreover – in high-stakes scenarios – using simpler, explainable models has been strongly advocated (Busuioc, 2021; de Laat, 2018).

Legal opacity

Frequently, however, the perceived lack of transparency in AI systems is not primarily due to their technical opacity. Instead, the core of the problem often stems from the opacity of AI systems provided by various legal mechanisms such as trade secret protection, business confidentiality, and non-disclosure agreements (NDAs), which are used to keep certain information hidden from external view (cf. Ananny & Crawford, 2018; Larsson & Heintz, 2020). As most AI providers choose to keep the inner workings of their software confidential[28], even simple models like decision trees can effectively become 'black-boxes' to external parties, including the regulators.

This reason for AI opacity is often referred to in the literature as *intentional opacity* (Burrell, 2016), or *legal opacity* (Tschider, 2021). Although AI providers may have legitimate reasons to protect sensitive information about their AI systems, much of the criticism focuses on applying confidentiality barriers as a 'form of self-protection by corporations' (Burrell, 2016), which is said to be frequently used to hide anticompetitive, discriminatory, or careless conduct behind a veil of inscrutability (Pasquale, 2015).

Some authors have argued that legal opacity of AI systems in some cases may challenge the rule of law (e.g. Fortes et al., 2022). The issue that is often

[27] For example, Zech et al. (2018) reported that an AI system trained using images from multiple hospitals to detect pneumonia from chest X-rays, has learned to associate specific hospital-specific markers, such as metal tokens present in the images, with the presence of pneumonia.

[28] There are very few algorithmic models available to freely access by means of open sourcing, yet even then most companies in AI domain keep parts of their code undisclosed to avoid abuse (Kemper & Kolkman, 2019).

pointed to in this context is the right to due process in ADM used in the public sector, when algorithmic opacity undermines the transparency of decision-making process. In criminal law, Taylor (2023) argues that 'a principle of "meaningful public control" should be met in all sentencing decisions if they are to retain their condemnatory status'. According to the author, officials who represent the public must remain morally accountable for sentencing decisions. While this approach does not rule out the use of algorithms in the public sector, it does require controls on how they are developed and applied in high-stakes areas. Perhaps the most well-known in this context is the US case Loomis v. Wisconsin, in which the risk assessment software sentenced Eric Loomis to six years in prison (see, for instance, Freeman, 2016). Yet, due to the trade secret protection, the algorithm remained a 'black-box' for the defendant, unavailable for the due process review.

While legal opacity of AI systems can in itself be seen as a challenge to the rule of law where such tools are used in ADMs by public authorities, researchers have also highlighted the negative implications of large-scale use of flawed AI tools which undermines fairness and exacerbates social inequalities. For example, Cathy O'Neil (2016) argues that, contrary to the belief that data and algorithms are neutral or objective, many of them are built on biased assumptions and data, leading to systematic discrimination and exclusion. Such systems – that O'Neil calls 'Weapons of Math Destruction' (WMDs) – are opaque, unregulated, operate at large scale, and often cause real-world harm, especially to disadvantaged groups.

Certainly, one might counter that these examples reveal problems not unique to AI, but already present in pre-existing social systems, which have long been marked by bias and unfairness. As Mulder et al. (2021) note, bias in judicial decisions has a long and well-documented history. From this point of view, algorithmic decision-making might in fact help *improve* fairness and equal treatment. However, technical or legal opacity of AI systems often prevents such embedded flaws to be identified and duly addressed. Transparency in automated decision making is therefore necessary to challenge such outcomes. This is essential not only for individuals – often from marginalised groups – directly affected by ADMs, but also for providing a way to ensure robustness and fairness of such systems on the societal level.

Although concerns about legal opacity of AI and rule-based ADM systems have traditionally focused on their use by public authorities, similar challenges are emerging in the private sector as well. For instance, automated decision-making is employed in credit and loan assessments, yet individuals often have little insight into how these decisions are made or how they can be contested.[29]

Often, however, risks associated with AI are difficult to identify due to their *high reach* – that is, their use on broad, societal scale. This is in particular challenging in contexts of most sophisticated and complex AI systems deployed by dominant technology companies operating globally.

High-reach AI systems

Transparency challenges are intensified by the scale at which dominant online platforms use ultra-personalised algorithms, often reaching billions of users. Among these, recommender systems — powered by machine learning and embedded in all major social media platforms — represent a particularly pervasive and influential form of high-reach AI (as explored in Paper II).

In the context of AI systems deployed at scale, even minor adjustments in algorithmic parameters may result in significant, cumulative effects, especially over time. High-reach recommender systems have been shown to be associated with a range of harms, including challenges to privacy and human autonomy (Yeung, 2017), amplification of disinformation (Celliers & Hattingh, 2020), and radicalisation tendencies (Hong & Kim, 2016; Ribeiro et al., 2020). Moreover, the companies collect far more data than needed to improve their services — what Zuboff (2019) calls 'behavioral surplus' — to better predict and influence user behaviour. As Zuboff argues, personal data that the largest technology companies gather, analyse, and profit from, is not only used to predict behaviour but increasingly to shape and control it, primarily for commercial gain.[30] At the same time, the internal workings

[29] This issue will be further discussed in Section 5.1.3

[30] As Zuboff (2019) puts it, the surveillance capitalism aims to 'reverse, subdue, impede, and even destroy the individual urge toward psychological self-determination and moral agency'(p.31).

of such systems, including the parameters which such algorithms are optimised for, are closely guarded secrets[31].

Moreover, the emergence of generative AI (GAI) platforms has added an additional layer of complexity to the transparency challenges in the digital environment. As Hagendorff (2024) observes, GAI raises a broad set of ethical concerns, including fairness (due to biased training data), safety and security (through harmful or malicious outputs), and privacy (via unintended data exposure). Other issues include *hallucinations* — i.e. producing plausible but false content, copyright and interaction risks, and value misalignment, where system goals deviate from human intent (Hagendorff, 2024). The combination of legal and technical opacity, in particular in contexts of high-reach of AI systems, significantly complicates efforts to identify and address such risks.

As will be further elaborated in the forthcoming section, the EU policymaking process has regarded the risks posed by AI systems — arising from their technical and legal opacity, as well as the high reach of certain AI systems — as crucial to address in order to ensure safe and trustworthy AI development and deployment within the Union.

2.2 The policymaking

EU policymakers have closely followed the rapid developments in the AI field and the growing interest of digital platforms in collecting personal data from European users. In general, EU policy development has been shaped by the ongoing technological changes, adjusting to newly identified risks as they arise.

Reviewing the Commission's policy documents shows a recurring pattern: on the one hand, it has aimed at encouraging innovation and uptake of AI

[31] For instance, digital platforms such as Facebook usually prohibit 'scraping' of data collected by independent researchers with the help of volunteer-installed browser extensions (Leerssen, 2021). As will be discussed further in this thesis – and in particular in Paper II – the DSA has introduced an avenue for researchers to examine large-scale risks. Nonetheless, the implementation of these provisions by many VLOPs remains limited.

technologies, and on the other hand, stressing the need to mitigate their potential negative consequences.

Although the concept of Trustworthy AI has become central in EU policymaking on AI technologies, it is important to recognise that relevant policy developments extend beyond AI-specific initiatives. Rather, they are embedded in the broader trajectory of the EU's digital strategy. This overview of EU policymaking provides a wider context for how AI governance has been envisioned and grounded in regulatory efforts around personal data and the digital environment.

Early stages of EU policymaking on digital technologies

One of the first important EU policy documents which clearly outlined the European direction for digital technologies was the adopted in 2010 EU Commission's *Digital Agenda for Europe* (European Commission, 2010). The document was primarily aimed at fostering a digital single market in the EU, 'to chart a course to maximise the social and economic potential of ICT, most notably the internet'. Importantly, the *Agenda* emphasised the urgent need for revision of the EU data protection law and evaluation of the rules governing online markets and services, with a view to 'enhancing individuals' confidence and strengthening their rights'. Overall, the EU's vision for digitalisation was presented in optimistic and forward-looking terms. We read, for instance, that:

> Wider deployment and more effective use of digital technologies will (…) enable Europe to address its key challenges and will provide Europeans with a better quality of life through, for example, better health care, safer and more efficient transport solutions, cleaner environment, new media opportunities and easier access to public services and cultural content.

To nurture such ambitious objectives, it was necessary to reform the data protection rules, which were at the time governed by the Data Protection Directive (95/46/EC) adopted in 1995[32]. As was envisioned by the EU Commission, 'the reform will first of all benefit individuals by strengthening

[32] As stated by the Commission in its 2012 Communication on *The Safeguarding Privacy in a Connected World: A European Data Protection Framework for the 21st Century* (European Commission, 2012b), the Data Protection Directive was adopted 'when the internet was in its infancy', while in 'today's new, challenging digital environment, existing rules provide neither the degree of harmonisation required, nor the necessary efficiency to ensure the right to personal data protection'.

their data protection rights and their trust in the digital environment' (European Commission, 2012b).

In 2015, the Commission presented the communication *A Digital Single Market Strategy for Europe* (European Commission, 2015), which set out initiatives aiming at providing better access to digital goods and services, and strengthening the potential of the EU digital economy. At that time, the impact of the platform economy on the personal data has been already clearly recognised. As the Communication stated, '[p]latforms generate, accumulate and control an enormous amount of data about their customers'. The Commission has therefore launched 'a comprehensive assessment of the role of platforms, including in the sharing economy, and of online intermediaries', covering issues such as transparency in search results and advertising, platforms' usage of the information they collect, and the ways of addressing the problem of illegal content on the Internet.

The EU's approach to AI technologies

In one of the first statements concerning AI in the EU policymaking – in the Commission's 2017 communication on the implementation of the Digital Single Market Strategy (European Commission, 2017) – AI has been framed as key technology bringing 'major benefits to our society' and 'a key driver for future economic and productivity growth'. The benefits of AI have been exemplified as leading to, for instance:

> (…) fewer fatalities on roads, smarter use of resources such as energy and water, less pesticide use on farms, and a more competitive manufacturing sector. In healthcare, robots already help with higher precision in surgery, among other tasks. They also assist in dangerous situations, for example in rescue operations following earthquakes or nuclear disasters (European Commission, 2017).

The same year, the EU Council stated that the EU needs to address with 'a sense of urgency' the emerging trends such as AI while at the same time 'ensuring a high level of data protection, digital rights and ethical standards' (European Council, 2017). The Council invited the Commission to propose a European approach to AI, and called on the Commission to put forth the necessary initiatives for 'strengthening the framework conditions with a view to enable the EU to explore new markets through risk-based radical innovations and to reaffirm the leading role of its industry' (European

Council, 2017). Consequently, in 2018, the EU policymaking on AI has been formally initiated.

The EU Commission's first steps in the AI policymaking were presented in April 2018 in the communication *Artificial Intelligence for Europe* (European Commission, 2018a). The Communication set out a European initiative on AI, stating that the 'EU should be ahead of technological developments in AI and ensure they are swiftly taken up across its economy.'

Building on this foundation, prepared together with the Member States, the Commission released the *Coordinated Plan on Artificial Intelligence* (European Commission, 2018b) in December 2018. The document outlined a strategic framework for national AI strategies, encouraged investments and collaboration among the EU Member States. Importantly, it also stressed the EU's commitment to human-centric AI: 'Europe can become a global leader in developing and using AI for good and promoting a human-centric approach and ethics-by-design principles.'

Moreover, the Communications were accompanied by the appointment of the High-Level Expert Group on Artificial Intelligence (AI HLEG). The task of the Group was to provide advice on investment strategies on AI and to develop AI ethics guidelines. The latter was meant to be developed together with all relevant stakeholders, with due regard to the Charter of Fundamental Rights of the EU, and with 'the ambition (…) to bring Europe's ethical approach to the global stage' (European Commission, 2018a).

The concept of Trustworthy AI

After broad consultations, the Ethics Guidelines for Trustworthy AI (AI HLEG, 2019) have been presented to the public in April 2019. The Guidelines are an important policy document, setting out a roadmap for ensuring and scaling Trustworthy AI in the EU. As the Guidelines state, their objective has been to 'make ethics a core pillar for developing a unique approach to AI, one that aims to benefit, empower and protect both individual human flourishing and the common good of society'. The Guidelines have served as a blueprint for AI development in the EU, and Member States were

encouraged to ground their national AI strategies on these shared principles (Larsson, 2021)[33].

In general, the framework presents AI technologies as a promising means to increase human flourishing, bringing progress and innovation, as well as enhancing individual and societal well-being, and the common good. The Guidelines incorporate the concept of *human-centric* AI, stating that AI systems need to rest on the commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom. This entails seeking to maximise the benefits of AI systems while at the same time preventing and minimising their risks.

Against this backdrop, the Guidelines present Trustworthy AI as a 'foundational ambition, since human beings and communities will only be able to have confidence in the technology's development and its applications when a clear and comprehensive framework for achieving its trustworthiness is in place.' Failing to live up to such commitments, as the Guidelines further explain, may result in 'preventing the realisation of the potentially vast social and economic benefits that they can bring.'

A succinct definition of what Trustworthy AI is has not been provided in the document. Instead, the Trustworthy AI is described in the framework as *lawful*, *ethical* and *robust*, which, as the Guidelines emphasise, should be met throughout the system's entire life cycle. Importantly, the first component – lawfulness – is only indicated in the Guidelines, and not developed further. It is, however, stressed and assumed that a trustworthy AI system is compliant with all applicable laws across the whole life-cycle of AI systems[34].

The main focus of the framework is on developing the latter two components – the ethicality and robustness of AI systems. With regard to the ethics component, the Guidelines point out that while many legal

[33] Trustworthy AI principles have been clearly reflected in, for instance, the AI Strategy in Poland, see: Söderlund, K. (2020).

[34] As stated in the Guidelines, these 'proceed on the assumption that all legal rights and obligations that apply to the processes and activities involved in developing, deploying and using AI systems remain mandatory and must be duly observed.' (AI HLEG, 2019)

obligations reflect ethical principles, adherence to ethical principles goes beyond formal compliance with existing laws.[35]

In turn, robustness of AI systems is explained in the Guidelines both from a technical perspective (ensuring the system's technical robustness as appropriate in a given context, such as the application domain or life cycle phase), and from a social perspective (in due consideration of the context and environment in which the system operates). Moreover, the Guidelines take a view that the responsibility for broader impact of AI systems should be taken not only for the intended but also unintended purposes, since 'even with good intentions, AI systems can cause unintentional harm' (AI HLEG, 2019). Thus, even if an ethical purpose is ensured, individuals and society must also be confident that AI systems will not cause any unintentional harm, and 'safeguards should be foreseen to prevent any unintended adverse effects of AI systems' (AI HLEG, 2019).

The aspects of ethicality and robustness of AI systems are further developed on three levels of abstraction – from the most high-level *principles*, more specific *requirements,* and an *assessment list*.

On the most general, *principle level*, the Guidelines stipulate that AI systems should be developed, deployed and used in a way that observes the ethical principles of *respect for human autonomy, prevention of harm, fairness* and *explicability.* The Guidelines point to the need for paying particular attention to impact of AI systems on vulnerable groups, such as children, persons with disabilities, historically disadvantaged or at risk of exclusion. The Guidelines further underline the need to consider the situations of informational power asymmetries between, for instance, businesses and consumers. Moreover, the framework points out that proportional and adequate mitigation measures should be adopted with respect to the risks posed to AI, in accordance with the magnitude of the risks.

On the second level of abstraction, the framework translates these ethical principles into seven key *requirements* that AI systems should implement and meet throughout their entire life cycle: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4)

[35] The Guidelines point to the need for referring to ethical principles to fill the regulatory gaps, since laws are 'not always up to speed with technological developments, can at times be out of step with ethical norms or may simply not be well suited to addressing certain issues.'

transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability.

The third level provides a non-exhaustive assessment list aimed at operationalising the trustworthy AI requirements. At the same time, as the Guidelines point out, the assessment should be tailored to the particular context, thus compliance with the Guidelines is 'not about ticking boxes', but about continuously identifying challenges, evaluating compliance with the requirements, implementing solutions.

The Guidelines thus provided an important and ambitious starting point for the discussions about 'Trustworthy AI for Europe'. Crucially for this thesis, the Guidelines have outlined the role of transparency in the governance framework. Interestingly, however, the Guidelines 'explain' this requirement not by providing any definition, but rather by referring to other concepts. It is stated that the transparency requirement 'is closely linked with the principle of explicability and encompasses transparency of elements relevant to an AI system: the data, the system and the business models' (AI HLEG, 2019). Further, the Guidelines point to other concepts – traceability, explainability and communication.

Traceability is explained as documenting the AI system's decisions, algorithms, data gathering and labelling, to enable of the auditability, explainability, and identification of the reasons of errors, to prevent future mistakes.

The concept of *explainability*, under the Guidelines, refers to 'the ability to explain both the technical processes of an AI system and the related human decisions', such as the application areas of a system. Technical explainability means that the decisions made by an AI system can be understood and traced by human beings (which seems to resemble the notion of traceability above). Further, it is noteworthy that the Guidelines point to the 'trade-offs' which might need to be made between the AI system's explainability and accuracy. What is also important to note is that the Guidelines highlight that '[w]henever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process', which echoes the wording of Art. 22 GDPR. The framework also points to the significance of 'explanations of the degree to which an AI system influences and shapes the organisational decision-making process,

design choices of the system, and the rationale for deploying it', thus indicating the broad understanding of 'AI transparency'.

The third concept – *communication* – means that users should be informed when they are interacting with an AI system. Moreover, in what seems to also allude to the GDPR's Art. 22, the Guidelines state that users should have an option to 'decide against this interaction in favour of human interaction', when the need to ensure compliance with fundamental rights arise. Furthermore, as the Guidelines explain, this principle should also mean that the AI system's capabilities and limitations are clearly communicated to 'AI practitioners or end-users' in an appropriate way, including the AI system's level of accuracy and limitations. This suggests that the transparency measures should be designed in a way that serves the transparency objectives of different stakeholders groups.

EU Commission's White Paper on AI

The next phase in EU policymaking was marked by the initiation of two major regulatory efforts aimed at governing digital and AI technologies. In the Commission's 2020 strategy *Shaping Europe's digital future* (European Commission, 2020b), one of the key actions launched concerned the overhaul of rules governing digital services as part of the Digital Services Act package. Second important initiative was the White Paper on Artificial Intelligence, which set out options for a legislative framework for Trustworthy AI.

Regarding the former initiative, the duality of objectives (embracing the potential while addressing risks) is displayed in the Explanatory Memorandum (European Commission, 2020a) to the Digital Services Act:

> Since the adoption of Directive 2000/31/EC1 (the 'e-Commerce Directive'), new and innovative information society (digital) services have emerged, changing the daily lives of Union citizens and shaping and transforming how they communicate, connect, consume and do business. (…) At the same time, the use of those services has also become the source of new risks and challenges, both for society as a whole and individuals using such services.

The European Parliament adopted resolutions (European Parliament, 2019b, 2019a), which included a strong call for maintaining the core principles of the e-Commerce Directive, protecting fundamental rights and online anonymity where possible, and ensuring transparency, accountability, and effective

obligations to tackle illegal content. They also called for strong public oversight and cross-border cooperation among authorities.

Building on the policy documents above, the Commission's White Paper *On Artificial Intelligence - A European approach to excellence and trust* (European Commission, 2020c) was presented in February 2020.

In the document, the Commission reiterated its commitment to 'a regulatory and investment oriented approach with the twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of this new technology'. The Commission has also highlighted its overall vision for the EU approach to AI, by stating that in addressing 'the opportunities and challenges of AI, the EU must act as one and define its own way, based on European values, to promote the development and deployment of AI.' Furthermore, the Commission emphasised that AI should benefit the whole EU population, by 'ensuring that new technologies are at the service of all Europeans – improving their lives while respecting their rights'.

Importantly, the White Paper endorsed the Ethics Guidelines for Trustworthy AI, and set out policy options on how to achieve these ambitious objectives. The White Paper was structured around the two main building blocks. The objective of the first one – 'ecosystem of excellence' – aimed at 'setting out measures to align efforts at European, national and regional level', in order to avoid duplications in research efforts within the EU. This would mean addressing the problem of fragmented landscape of centres of competence, to mobilise resources in research and innovation, and to create the right incentives to accelerate the adoption of solutions based on AI. The Commission stipulated that the goal is to generate at least over €20 billion in AI-related investments per year within the EU throughout the next decade (European Commission, 2020c)

The second building block – 'ecosystem of trust' – focuses on the regulatory framework for AI. The framework is based on the assumption that 'it must ensure compliance with EU rules, including the rules protecting fundamental rights and consumers' rights, in particular for AI systems operated in the EU that pose a high risk'. The stated objective of the framework was to ensure that EU citizens feel sufficiently confident to

engage with AI applications, and that businesses should have 'the legal certainty to innovate using AI'.

Notably, the Commission highlighted the regulatory gaps in AI governance within the EU. Firstly, the White Paper pointed to the enforcement challenges stemming from the lack of transparency in AI, making it difficult to identify and investigate legal breaches. Secondly, the document indicated that existing product safety laws mainly cover tangible goods but do not clearly regulate stand-alone software or AI-driven services. Importantly, the general EU safety legislation at that time applied to products and not to services, and 'therefore in principle not to services based on AI technology either'. Thirdly, the Commission recognised that AI's changing functionality — such as self-learning and software updates — introduces risks that the safety frameworks did not sufficiently address. Fourthly, the problem of responsibility allocation would arise when AI components are added to products by third parties, complicating liability rules. Finally, the Commission highlighted the need to revise the safety frameworks, in view of the new risks such as cybersecurity threats and connectivity failures, which AI introduces.

The White Paper concludes that the EU legal framework may need new legislation specifically addressing AI to keep up with technological and commercial developments. The proposed regulatory approach would focus on maintaining a balance between effectiveness and avoiding excessive burdens. To this end, the Commission proposed a risk-based approach for AI regulation, based on the sector of AI applications and their intended use. High-risk AI applications would be subject to stricter requirements to safeguard safety, consumer rights, and fundamental rights, and the projected requirements would include such areas as training data, record-keeping, transparency, robustness, accuracy, human oversight, and specific provisions for applications like remote biometric identification.

The ideas presented in the EU policymaking above have since been translated into binding legal obligations across several EU regulations. This thesis focuses on three of them – the GDPR, DSA and AIA – which will be briefly outlined below.

2.3 The law

As shown above, the regulatory frameworks analysed in this thesis are products of many years of discussions and agenda-setting within the EU's policymaking processes. The EU regulations that are in focus in this thesis address different aspects of AI governance. First, I present the GDPR framework, as it sets out the legal framework for personal data processing — a foundation relevant to many AI systems. The second framework outlined here is the DSA, in which my focus has been on *high-reach* AI systems. Third framework examined in this thesis is the AIA, which specifically regulates the use of AI technologies in the EU, and which focuses on prohibited and high-risk uses that may undermine fundamental rights and other important EU principles.

2.3.1 The General Data Protection Regulation (GDPR)

The General Data Protection Regulation (GDPR) (Regulation 2016/679), was adopted as an update to the mentioned earlier Data Protection Directive (95/46/EC). The GDPR has been applicable from 2018, and has important implications on the AI uses. The GDPR is a technology-neutral regulation, meaning that its provisions apply regardless of the specific choice of technology used (AI, ADM, or other automated technologies), or whether the data is processed manually. The GDPR is one of the most comprehensive data protection frameworks in the world, and applies to any organisation that processes the personal data of EU citizens.

The Regulation establishes rules for the collection, storage, and other forms of personal data processing, and was designed to enhance the protection of individuals' personal data. For example, individuals have the right to access their data, request corrections, and ask for data deletion. The GDPR imposes certain obligations on organisations, including those on transparency and accountability. Non-compliance with the GDPR can result in significant penalties, with fines reaching up to 4% of the global annual turnover[36].

[36] However, the actual use of such enforcement provisions by the national oversight bodies will be further discussed in Section 5.2.

The provisions of the GDPR that are in particular relevant in this thesis are the transparency measures which are required from data controller vis-à-vis data subjects and oversight bodies. These will be presented in more detail in Section 5.2.

2.3.2 The Digital Services Act (DSA)

The Digital Services Act (Regulation 2022/2065) was adopted in November 2022 as an update of the E-Commerce Directive (Directive 2000/31/EC), modernising the governance framework for digital services in the EU. The DSA is relevant for AI governance, as various forms of AI are used by online platforms, such as recommender systems, to suggest content to users and in content moderation. The territorial scope of application of the DSA covers, similarly as the GDPR, companies operating within the EU, regardless of where their headquarters are situated.

The Regulation has been fully applicable from February 2024, and applies to all intermediary services, such as internet providers, hosting services, and online platforms. The stated goal of the DSA is to increase accountability and transparency in the digital environment (European Commission, 2021). To this end, the DSA obliges the digital services operators to tackle illegal content, improve user safety, and provides users with additional rights to challenge content moderation decisions. Failure to comply with the DSA can result in substantial fines, up to 6% of a company's global annual revenue.

Although the DSA applies to all intermediary services, the most stringent transparency rules are imposed on very large online platforms (VLOPs) and very large online search engines (VLOSEs), with at least 45 million average EU users per month. The online providers which have been captured by the scope of the social media VLOPs include such companies as Facebook, Instagram, LinkedIn, Pinterest, Snapchat, TikTok, X (formerly Twitter), YouTube, and a range of platforms with pornographic content (European Commission, 2023).

Among the transparency requirements for the VLOPs, in addition to other duties to moderate content and combat illegal content, is to mitigate the negative impact of their algorithms on societies. These include, for example, the requirement that VLOPs should 'identify, analyse and assess any

systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services', such as the dissemination of illegal content through their services, any actual or foreseeable negative effects for the exercise of fundamental rights[37], effects on civic discourse and electoral processes, and public security, effects in relation to gender-based violence, the protection of public health and minors, and serious negative consequences to the individuals' physical and mental well-being.

The requirements above should take into account the design of the recommender systems (as defined in Art. 2 DSA) and any other relevant algorithmic system, such as systems for selecting and presenting advertisements, the amplification and potentially rapid and wide dissemination of illegal content.

The Commission is responsible to enforce the DSA together with national authorities, who supervise the legal compliance of the platforms established within their territories. The Commission is primarily responsible for the monitoring and enforcement of the additional obligations applying to VLOPs and VLOSEs, such as the measures to mitigate systemic risks.

2.3.3 The AI Act (AIA)

The third piece of EU legislation analysed in this thesis is the Artificial Intelligence Act (AIA) (Regulation 2024/1689), which can be seen as a direct product of the EU policymaking on the Trustworthy AI[38]. Adopted in 2024 and applicable from 2025, this Regulation has been introduced into the EU legal framework primarily to address the most significant risks arising from AI technologies. As mentioned earlier, the AI Act includes a definition of what counts as AI systems for the regulatory purposes.

---

[37] These include, in particular, the right to human dignity, respect for private and family life, the protection of personal data, freedom of expression and information, non- discrimination, respect for the rights of the child, and to a high-level of consumer protection.

[38] Notably, the stated purpose of the AIA is 'to improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI)', thus the Trustworthy AI aim has been incorporated in Art. 1 AIA.

AI Act is a risk-based regulation, meaning that the level of legal requirements depends on the potential risks that the AI systems pose on fundamental rights and human safety. AI systems are categorized into four levels: unacceptable risk (banned), high risk (strictly regulated), limited risk (subject to certain transparency requirements), and minimal risk (little to no regulation). High-risk AI applications — such as those used in biometrics, employment, law enforcement, or medical devices — must meet certain standards, including those on transparency, data governance, accountability, and human oversight.

Generative AI, a type of AI technologies which emerged while AI Act was negotiated, is also covered by certain rules. These include, for instance, requiring certain information on training data, potential biases, and synthetic content detection.

Just as in case of the two frameworks above, AI Act applies to all AI systems operating or developed in the EU. The penalties for AIA violations have been set even higher than for the GDPR and DSA, for up to 7% of the global annual turnover (Art. 99 AIA).

In sum, the present chapter has aimed to outline the broader technological, policymaking, and regulatory context in which the concept of AI transparency is situated. I have briefly presented the evolution of AI, its risks, the corresponding EU policy responses, and how these have shaped the development of the binding legal instruments — GDPR, DSA, and AIA. These developments reflect the growing role of transparency as a core governance tool for addressing AI risks, promoting Trustworthy AI, and ensuring alignment with EU laws and values.

In the following chapter, I move from context to concept, exploring transparency as operating in our language in many ways, and which is by no means limited to its legal expressions. Drawing on scholarly literature, I introduce various theoretical perspectives that help unpack the conceptual foundations of transparency, its underlying meanings and limitations.

3 Conceptual perspectives on transparency: from literal meaning to 'floating signifier'

As the previous chapter illustrated, transparency plays an important role in AI governance frameworks. Before turning to the specific notion of 'AI transparency', this chapter introduces some of the conceptual perspectives concerning the concept of transparency in general. As will be further elaborated in Chapter 5, much of the conceptual entanglements of the term 'AI transparency' arise because the concept of transparency is far from being transparent itself.

Indeed, despite the increasing popularity transparency – or perhaps because of it – the exact meaning of the term 'transparency' is difficult to capture in one, all-encompassing definition. In the sections that follow, I explore some of the ways in which transparency is used in our language: its literal definition, its metaphorical use, its role as a governance ideal, its translations as governance tools, and how it has come to function as an 'ideograph' that 'can be filled in different strategic ways' (Meijer, 2014), or – what I describe here as – a 'floating signifier'.

3.1 From literal meaning to metaphor

The etymology of the term 'transparency' can be traced to the Latin *trans* ('through, beyond') and *parere* ('to appear, to come in sight') [39]. In its most literal sense, therefore, it describes a physical characteristic — a material's capacity to let light pass through, allowing visibility of what is behind. As Koivisto (2022) puts it, it is an optical condition of a 'medium' through which other objects become visible.

[39] Oxford English Dictionary explains transparency as the 'quality or condition of being transparent, perviousness of light, diaphaneity', and Cambridge Dictionary describes it as 'the characteristic of being easy to see through'.

The term ‘transparency’ is still, at times, used in its original sense to describe the physical visibility through an object. However, transparency is nowadays used in our language in a multitude of ways, and these are primarily built on its meaning as a *metaphor*. These metaphorical uses expand the term to encompass ideas like accessibility to knowledge, openness, publicity, understanding, as well as in social relations – as a value attributed to other people (Koivisto, 2022). And, as mentioned, the extent to which transparency as a metaphor has been used across domains have in some cases rendered transparency to function as a ‘floating signifier’, devoid of specific meaning.

To better understand why *transparency* has become such a widely used term in today’s governance discourses, it would be useful to consider how metaphors are formed in general. Reddy (1979), for example, describes metaphors as a way of mapping physical objects onto mental processes. In essence, metaphors operate across two levels of abstraction. As Koivisto (2022) explains, this involves linking a *source domain* — a more concrete phenomenon — with a *target domain* — a more abstract concept. Certain attributes from the source domain are then transferred to the target domain by analogy. In case of the transparency metaphor, the optical property of an object (the source domain) becomes associated with abstract ideas such as human character, governance, or regulatory mechanisms (the target domain).

An important point regarding transparency as a metaphor is that it extends beyond mere visibility and is often associated with knowledge and understanding. Commonly, different expressions which are based on vision or light relate to our ability to *understand*. For instance, we ‘clarify’, ‘highlight’, ‘elucidate’, which creates in our minds the impression that by seeing a phenomenon, we may understand it (the ‘knowing is seeing’ metaphor) (Koivisto, 2022). As Koivisto, (2022) observes, metaphors are not merely decorative speech. Rather, they tap into fundamental cognitive patterns shaped by our embodied experience of the world.

The role of metaphors in our communication is therefore more powerful than it may initially seem. As Calo (2016) points out, a metaphor connects unlike concepts to achieve rhetorical impact, thus ‘every metaphor is, in its own way, an argument.’ Metaphors have, therefore, the power to clarify

our understanding while also shaping — or even distorting — how we perceive things (Solove, 2024).

3.2 The metaphor of transparency as a governance ideal

As seen above, transparency operates in our language not only in its literal meaning – an optical condition, but also as a suggestive metaphor, which may be accommodated in many contexts. In this section, I describe how transparency as metaphor has become a symbol of the governance ideal, serving as an illustration of the well-functioning accountability mechanisms in public and private governance.

Transparency as a concept used in the governance contexts has been clearly on the rise during the past three decades. Carolyn Ball (2009) explores how the metaphorical use of 'transparency' gained traction in the 1990s, particularly through anti-corruption initiatives by non-governmental and supranational organisations. Forssbæck & Oxelheim (2014) similarly observe that transparency became a widely used term in economic and political discussions, especially in response to financial instability in the mid-1990s and corporate scandals in the early 2000s. This increasing relevance is also visible in the growing number of transparency-related studies (Larsson & Heintz, 2020).

While transparency is gaining increasing attention in economic, organisational, and political contexts today, does this mean that it is a recent invention? Although its role in public governance appears to be a relatively modern development, the idea of transparency has much deeper roots, with origins traceable to philosophical and religious traditions throughout history.

3.2.1 Philosophical roots of transparency

Transparency, in its metaphorical sense, carries with it philosophical associations that date as far back as to antiquity, including Plato's and Aristotle's understandings of truth and the purity of the soul (Koivisto,

2022). In the Middle Ages and the Christian theology, the rhetoric of light would refer to divine qualities of virtue and innocence, while darkness would be associated with sin and the diabolical (Koivisto, 2022).

The metaphor of transparency and its associations with light and knowledge later found fertile ground in the Enlightenment, a period marked by the pursuit of clarity and rational understanding. Enlightenment is often seen as the time when the ideological roots of transparency in its modern meaning have been established, although the most prominent philosophers of that time, such as Jean Jacques Rousseau, Jonathan Bentham and Jean Bodin, did not use the term 'transparency'. Busuioc et al. (2023) observes that transparency was then understood as the right to *publicity* and the *access of the public to the documents*. Bentham referred to publicity as a way to expose what was inaccessible to public scrutiny, thus it could be seen as including the same metaphorical dimension that is now vested in transparency (Meijer, 2014).

Another idea preceding transparency in the Enlightenment was *openness*, which also exists in the modern, transparency-related terminology. In Bentham's works, the concept of openness was closely linked to the idea of the open government, as a way to prevent abuses of state power, which has also had a strong impact on the development of the modern public sector (Meijer, 2014). As Meijer (2014) writes, '[t]he most important transparency measures, such as opening up of archives, public sessions of representative bodies, and the publication of government documents, can all be traced back to Bentham's ideas on openness.' The notion that people behave correctly when they are being watched can also be linked with Bentham's idea of *panopticon*.[40]

In a similar vein, Christopher Hood (2006) observes that French revolutionaries embraced the vision of a transparent society as one in which there was no room for the kind of social obscurity they believed gave rise to injustice and unhappiness. He identifies Rousseau as a central figure in this tradition. For Rousseau, transparency symbolised a return to the original

[40] The panopticon refers to an idea of a prison where inmates are under continuous observation from a central watchtower, with their constant visibility expected to promote compliance with rules. Although Bentham developed this concept for the prison context, the underlying principle—that transparency encourages better conduct—has been extended to the behaviour of public officials and politicians.

state of nature — a kind of moral restoration and as a longing for paradise, when Adam and Eve were fully 'transparent'.

However, not all philosophers in the 18th century argued in favour of transparency. Jean Bodin, for instance, defended the secrecy of the imperial policy and stressed that the King's ability to maintain the integrity of the state would be undermined by transparency (Meijer, 2014). Even though Rousseau's vision of a transparent society has become a dominant paradigm, particularly following the French Revolution, the debate between openness and secrecy continue to inform how transparency is understood in governance discussions today.

Given the rich philosophical heritage tied to the idea of transparency, it is worth reflecting on why related terms — such as openness, publicity, or access to documents – have been largely subsumed and even supplanted by the term 'transparency' in the modern-day debates. Koivisto (2022) suggests that the word holds a unique appeal that surpasses terms like *publicness*, *publicity*, *the right to know*, *access to knowledge*, *freedom of information*, and *openness*. This appeal, according to Koivisto, lies in the metaphor's strong visual quality and its ability to carry a wide range of positive associations — such as clarity, consistency, sincerity, purity, truthfulness, and efficiency. Similarly, Baume and Papadopoulos (2018) argue that transparency has overtaken related concepts precisely because it is 'better equipped' metaphorically, drawing on broader lexical fields than its conceptual counterparts.

3.2.2 Transparency as a concept embedded in law

As transparency has evolved from a suggestive metaphor into a governance ideal, it has also become embedded in legal frameworks — particularly within public administration, where it supports principles such as rule-based governance and democratic accountability (Koivisto, 2022). Transparency has long played a central role in the context of state governance, underpinning notions of public control and institutional legitimacy. As a socio-legal governance ideal, transparency is intertwined with asymmetrical power and its legitimatised use that is characteristic of the state and other

forms of public authority. It can thus be understood as a legal concept that draws on the Enlightenment roots discussed above as well (Koivisto, 2022).

It is worth noting that in the legal vocabulary the concept of transparency seems to be also on the rise. As Schudson (2015) observes, the term 'transparency' has become the preferred term only over the last couple of decades. Before that, in the legislative and administrative practices the term 'disclosure' was preferred, or other concepts such as 'access to information', 'open government', 'the principle of openness', 'the right to know', 'publicity', 'freedom of information', or even 'sunshine acts'.

The concept of transparency as a metaphor for governance ideal in public law and policy is discussed by Fenster (2010) in his article 'Seeing the State: Transparency as Metaphor'. The author argues that the idea of transparency in state governance plays out in two dimensions: the *metaphorical promise* ('the democratic wish' or 'aspirational goal') and its *technocratic translations*. Thus, as Fenster maintains, transparency in state governance operates as a theoretical concept, which can, at least in theory, be implemented in legislation (as various legal tools). This duality of transparency as operating on different levels of abstraction has been also adopted in this thesis, and transparency understood as technocratic translations of the 'democratic wish' will be further elaborated in Section 3.3 below.

Apart from its well-established presence in state and administration governance, transparency as a governance ideal in law is also a fundamental component of EU law, with transparency being an explicit governance choice made in the Lisbon Treaty and the Charter of Fundamental Rights. As Leino-Sandberg (2025) observes, the concept of transparency in EU law, among other contexts, places an obligation on the EU institutions to carry out their work as openly as possible, make decisions with openness, and ensure that legislative documents are published, with a view of guaranteeing that every EU citizen has the right to participate in democratic processes of the Union. To this end, public access rules have been laid down in Regulation (EU) 1049/2001 regarding public access to European Parliament, Council and Commission documents, to operationalise these principles.

3.2.3 Transparency as a precondition for accountability

Given the apparent prominence of transparency in legal and regulatory contexts, it is worth asking: why is transparency — or related concepts — so often framed as a governance ideal? It could be argued that the reason for it lies not in the value of transparency itself, but in other objectives that it *enables*. Indeed, Turilli and Floridi (2009) observe that transparency is not an ethical principle in itself, but is instead a *pro-ethical condition* that allows other principles to be realised. Thus, although transparency understood as a governance ideal is a central value, as such, it should primarily be seen as a means to achieve other objectives.

Although other objectives that transparency may contribute to are such values as democracy, answerability, responsiveness, and legitimacy[41], most commonly the need for transparency is justified by accountability objectives (Busuioc et al., 2023; Meijer, 2014). As Koene et al., (2019) highlight, transparency is *implied* in accountability, since 'if we cannot know what an organisation is doing, we cannot hold it accountable, and cannot regulate it'. In contexts of state power, Koivisto (2022) stresses the function of transparency as making unilateral or otherwise unequal power visible and as such, controllable. In in other words, those in power – the agents – should be accountable to the principals – to those from whom the power emanates[42].

While transparency is often associated with enhanced accountability mechanisms, it is important to recognise that information disclosure is not always constructive. In certain cases, as will be shown, transparency may even work against accountability rather than supporting it.

Indeed, transparency is conceptually closely linked to accountability, yet it differs from it on important accounts. As Busuioc et al. (2023) writes, while accountability is usually associated with control or holding power to account, transparency's aim is merely to render visible what has been

[41] However, emphasis should be put on the word 'may' – the issue of the *perverse effects* of transparency on accountability, legitimacy or democracy have been broadly discussed in the literature (see, for instance, Hood, 2007, 2010; Meijer, 2014).

[42] In this sense, transparency is understood as following the *democratic rationality*, which will be further explained in Section 3.3.2.

hidden, which may open up possibilities for accountability, control, or oversight, that otherwise would not be possible.

However, the link between transparency and accountability is not always straightforward. Meijer (2014) in an analysis of the relation between transparency and accountability proposes three routes between these concepts:

- *direct route: transparency as facilitating horizontal accountability* – when increased transparency contributes to accountability by providing citizens, stakeholders, and media with better access to information.
- *indirect route: strengthening vertical accountability* – the accountability forum is warned by a third party as soon as something untoward is observed in the conduct of a public official or a public organization. Access to information enables citizens and other stakeholders to contribute by acting as a 'fire alarm' for formal accountability forums.
- *inverse relation: transparency reduces the need for accountability* – transparency may also diminish the need for formal accountability mechanisms. In this line of argument, transparency is an instrument to ensure that actors conform to public standards that reduces the need for another instrument – accountability – to achieve this objective. This relation could be linked to Bentham's idea of *panopticon*.

The above links between transparency and accountability show transparency as having positive effects on governance mechanisms. However, the relations between the concepts may also be ambiguous or adverse. In this context, Hood (2010) illustrates the complex relationship between transparency and accountability through three metaphors: 'Siamese twins', indicating that the concepts are inseparable; 'matching parts', suggesting they work in harmony to support one another; and 'an awkward couple', pointing to the potential for friction when they are implemented together.

Such varying relational framings reveal that transparency is not always inherently or universally beneficial; rather, its effects are context-dependent and may entail trade-offs. As will be further explained in the

following section, while transparency can enhance accountability, support democratic oversight, and improve institutional trust, it can also introduce tensions — particularly when it collides with other legitimate values or objectives.

3.2.4 The interplay between transparency and secrecy

In governance settings — and even when transparency is approached as a governance ideal — it is essential to consider the role of its conceptual opposite: secrecy. Why this is important? Transparency as a governance ideal must often be weighed against competing interests, such as the need to protect privacy and other rights of individuals, as well as other types of sensitive information, including business confidentiality, trade secrecy, or national security. This leads to a broader and more nuanced consideration of transparency's role as a governance ideal – it does not exist in a vacuum but is always negotiated in relation to the boundaries set by legal, ethical, and political constraints.

Thus, although secrecy is often intuitively perceived as a negative value (see e.g. Koivisto, 2022), it frequently serves a legitimate function. Within governance frameworks, such legitimate interests must be weighed against one another, and in many situations, the justification for maintaining secrecy can be as compelling as that for ensuring transparency. However, striking the right balance between these competing interests is context-dependent and rarely straightforward. Legal, cultural, and societal norms influence how this balance is negotiated — and as these norms shift, so too do the weightings of different values. Transparency, then, is not a fixed standard, but an evolving ideal — continuously shaped by public demands, legislative choices, judicial interpretations, and broader societal and technological change.

In the history of political thought, as mentioned earlier, philosophers such as Jean Bodin argued that secrecy in public governance is central to maintain the integrity of the state. Indeed, transparency understood as information disclosure may in many cases do more damage than good. One could say that the proponents of transparency see it a panacea for all kinds of ills, while opponents see it as having a negative effect on democracy

mechanisms (Meijer, 2014). Other authors point out the negative effects of transparency, since uncontrolled information disclosure may undermine public trust. Onora O'Neill (2002), for instance, argues that 'trust seemingly has receded as transparency has advanced'. This aligns with the argument of Koivisto (2022) that only intentional, controlled public governance transparency may positively contribute to legitimation of power.

Moreover, O'Neill argues that a flood of unsorted information may lead to more uncertainty and may, as a result, confuse accountability. In such cases, transparency may overwhelm rather than inform, resulting in decreased trust. She warns that the unchecked expansion of transparency can foster a 'culture of suspicion'. Seen in this light, after all, transparency may not be a panacea to all wrong ('the more information – the better'). Rather, its positive effects depend on the way transparency is designed and implemented in governance frameworks.

For example, freedom of information (FOI) legislation does not grant unrestricted access to all types of information. Instead, it establishes the principles and boundaries for accessing public documents, including rules for disclosure and legally defined limitations (see e.g.Olsen et al., 2024).[43] In the European Nordic countries, especially in Sweden, there has been a long tradition of openness of public documents (*offentlighetsprincip*), which Julia Björverud (2024) describes as being in a continuous dialogue with the principle of secrecy (*sekretess*). This relationship and dynamics play out differently in national (or EU) legal systems. In Sweden, for instance, the starting point is openness, with secrecy treated as an exception. These exceptions are narrowly defined by law and justified on grounds such as individual privacy, national security, or intellectual property rights. Following the Swedish Freedom of the Press Act (Tryckfrihetsförordning [1949:105]), access to public documents may only be restricted when necessary, for example to protect national security, oversight activities, crime prevention, or personal and economic privacy (see Chapter 2 of the Act).

In contrast, it should be noted that AI governance frameworks generally take secrecy as their starting point, as most AI systems are developed by

[43] The Freedom of Information (FOI) legislation will be further explained in Section 3.3 as an example of transparency tools.

private companies. This becomes particularly contentious when such systems are used in the public sector, where FOI obligations still apply (cf. Olsen et al., 2024).

This tension between openness and secrecy illustrates the complex and often contested role of transparency in contemporary governance discourses. Important questions arise not only about access to information but also about the design of transparency itself: who transparency is for, what form it takes, and how it can be made actionable within regulatory frameworks. These questions signal a shift from transparency as an abstract principle to transparency as a set of operational tools embedded in governance practice, which will be explored in the following section.

3.3 The metaphor of transparency as a governance tool

In the preceding section, transparency was presented as a governance ideal — a concept that operates at the level of abstract principles and competing interests. This section shifts focus to the *technocratic translations* of this ideal (Fenster, 2010). Here, transparency is no longer simply an aspirational metaphor but becomes a concrete regulatory tool. This transformation involves embedding transparency into legal instruments, operational mechanisms, and institutional practices.

3.3.1 Legal translations of transparency

As Koivisto (2022) observes, within the realm of law, 'transparency functions both as a governance ideal and as a governance tool, serving the overarching objectives of democracy and rule-governed administration' (p. 117). As has been pointed to above, transparency in law is subject to the interplay between secrecy and transparency interests. When all the relevant interests, principles, and values are weighed against each other during the legislative processes, it is hardly surprising that this is the point at which decisive outcomes emerge. Although a general consensus among stakeholders may appear to exist during the public discourse and initial

policymaking stages, the process of translating these ideas into concrete legal provisions frequently exposes underlying tensions. As the specific shape of rights and interests are negotiated, conflicting priorities become apparent, hidden agendas come to light, and some stakeholder groups may end up being dissatisfied with the resulting trade-offs[44].

Still, such transparency tools are the ones that ultimately the target actors must obey, and can be held accountable for. As such, transparency in the form of governance tools may appear as more tangible and observable expressions of the concept than as an abstract, governance ideal.

While transparency in law as various legal tools may take many forms, Koivisto (2022) suggests a useful way to categorise them – though the author points out that this is not an exhaustive classification of all possible technocratic expressions of transparency. Following Koivisto's categorisation of transparency as governance tools, the most prominent technocratic translation of transparency in law are the various forms of *documentation*, on which governance transparency greatly relies. Documentation is a necessary precondition of *access to documents,* which allows people to access representations of governmental power (Koivisto, 2022). As such, transparency in form of access to publicly held documents by citizens is primarily realised by the mentioned above freedom of information (FOI) rules. These transparency measures have become basic requirements of democratic governance, and they are often even regarded as a self-evident good in society (Etzioni, 2010).

The second group of governance tools refers to the *attendance of the public*, which allows citizens to witness governmental decision-making in action, providing oversight beyond written records (Koivisto, 2022). Also in broader governance contexts, it enables citizens, journalists, and other observers to gather contextual cues — such as tone, hesitation, or informal interactions — that are often absent in written documents. In this sense, public attendance becomes not only a means of scrutiny but also a mechanism for building trust, where transparency is enacted through visibility, proximity, and participation.

[44] This act could be summarised by a quote attributed to Otto von Bismark: 'Laws are like sausages – it's better not to see them being made.'

Communication, in turn, can be explained as a proactive approach to transparency, where governments shape their public image while also making governance more understandable and accessible (Koivisto, 2022). This form of transparency involves the strategic presentation of information, narratives, and institutional identities through various channels — press releases, official websites, or direct communication with the addressees. As such, transparency is not merely about access to documents or records, but about constructing visibility in a way that is intelligible, curated, and often persuasive. In doing so, communication functions not only as a tool for informing the public, but also reinforcing trust, and managing public expectations.

Moreover, while the above forms of transparency may be seen as *controlled forms* of transparency, as Koivisto (2022) points out, transparency can also manifest in *uncontrolled forms*, through unauthorized disclosures such as *leaks* and *whistleblowing*. Although these expressions of transparency appear in legal discourse, they typically fall outside the scope of transparency as a normative governance tool. By revealing confidential information held by governments or private actors, such actions are often in conflict with legal norms that define what should remain undisclosed (Koivisto, 2022).

3.3.2 The transparency directions

To better understand how transparency as governance tools function within formal governance structures, it is also useful to consider how the operation of transparency tools can be systematised. Following David Heald (2006), these can be sorted in terms of the direction of information flow[45]. The framework distinguishes between vertical and horizontal directions of transparency. Vertical transparency comprises two opposing flows: *upward transparency*, which is directed from subordinate actors or institutions toward a hierarchically superior body, and *downward transparency*, which flows from authorities toward those they govern. As Fox (2007) also observes, upward transparency at its extreme may take the form of state

[45] This framework was used to analyse the changes in the required information from large online platforms (VLOPs) that the DSA has required, in Paper II of this thesis.

surveillance. Conversely, downward transparency aligns more closely with democratic ideals, allowing members of the public to access information about governing bodies and their actions (Fox, 2007).

In contrast, horizontal transparency concerns relations between actors or institutions situated at the same hierarchical level. It includes *inward transparency*, where actors external to a system are allowed to 'view in', gaining access to internal processes or information, and *outward transparency*, where those inside a system are able to observe and interpret their external environment. These horizontal directions are particularly relevant in contexts such as organisational accountability, inter-institutional relations, or between consumers and businesses.

According to Heald, the ideal condition would involve a balanced presence of all four directions, resulting in what he refers to as *fully symmetric transparency*. At the opposite end of the spectrum lies *fully symmetric non-transparency*, where none of the directional flows are present. However, in practice, transparency is rarely evenly distributed. Instead, it is often asymmetric, with certain directions privileged over others depending on institutional design, political interests, or power relations (Heald, 2006).

In a conceptualisation of transparency directions that complements Heald's typology, Koivisto (2022) discerns two ways in which transparency as governance tools may operate – each rooted in a different governor-governed relationship.

The first is the *democratic rationality of transparency*, where the fundamental premise is that the governed are the ultimate holders of political power. In this configuration, transparency is aimed to scrutinise those who govern on behalf of citizens. As Koivisto explains, legitimacy in this model is derived from the idea that the governor's authority depends on the continued acceptance of the governed. The agent (the governor) is thus not only acting in the name of the principals (the governed), but is also accountable to them for the use of that power. Transparency here functions as a condition for democratic legitimacy, making visible how delegated authority is exercised and enabling contestation if needed.

The second model, referred to as the *public law rationality of transparency*, reflects the inverse configuration. In this case, the governor is the holder of legitimate political authority, authorised — legally and institutionally — to

regulate and oversee the conduct of the governed. Here, transparency serves as a mechanism through which the government ensures that individuals or companies comply with rules, norms, or standards.

The conceptual frameworks proposed by both Koivisto (2022) and Heald (2006) will be further developed and brought into dialogue in Chapter 6, where the various directions and dimensions of AI transparency are discussed.

3.4 From metaphor to 'floating signifier'

So far, the concept of transparency has been presented as either a visual metaphor, a governance ideal, and its legal translations in form of various governance tools. In this section, I extend the analysis by framing transparency as a *floating signifier*[46]. This characterisation aligns with that of Busuioc et al. (2023), who describe transparency as a 'floating signifier' – 'a malleable concept that is empty of specific content but rather refers to form'. The rapid and widespread adoption of transparency within governance discourses, as discussed earlier, may have contributed to this semantic drift.

This trend has been noted by other scholars as well. Meijer (2014), for example, observes that '[w]hen used in political debates, the term "transparency" is often not defined and kept ambiguous'. He refers to transparency taken in this way as an *ideograph* – 'something nobody can be opposed to but that is conceptually empty and can be filled in different strategic ways'. Scholtes (2012) highlights that the ambiguity of the concept of transparency makes it attractive to politicians, since transparency can be used to underpin a broad variety of political arguments. The author shows how transparency can be connected to democratic values, administrative control, public accountability, the promotion of market forces, and an attitude of openness.

[46] Oxford Dictionary describes the term as 'signifier without a specific signified'. See also Malabou (2022), who defines the term 'floating signifier' as 'depending on the context in which they appear and the relationships they form with other words at a given moment' (p. 255).

Koivisto's (2022) refers to this phenomenon as transparency *discursification*. As the author observes, it can be debated whether the phenomenon of transparency reaching in some contexts the status of a 'floating signifier' – or a discursified concept – is a negative or positive trend. On the one hand, this shift may reflect a widespread support for transparency as a core value in organisational governance. On the other hand, however, the positive associations of transparency as a suggestive metaphor may lead to its use in ways stripped down of genuine accountability mechanisms (Koivisto, 2022). In other words, if transparency frameworks are not supported by appropriate legal or other accountability mechanisms, the concept risks being watered down as a governance ideal. As Koivisto argues, using the concept of transparency in this uprooted form utilizes the positive qualities the concept represents, while avoiding accountability burdens. When used in this way, the positive aura surrounding transparency is often used intentionally as a rhetorical, *legitimising strategy*.

Moreover, Koivisto points to deeper causes of this trend. As she observes, the undercurrent that has driven the discursification of transparency has been the emergence of powerful actors in the private sector. In particular, this concerns large technology firms and dominant online platforms, which created a new, 'non-democratic power relation and the informational asymmetry'. These actors often invoke the transparency claims for the purposes of legitimisation of their practices. Yet, as they have emerged outside of the bounds of public governance structures, they are not subject to democratic accountability mechanisms as the public institutions are.

In view of the above observations that transparency increasingly operates as a 'floating signifier', it becomes even more important that transparency has its bearing as a governance tool. While in many contexts transparency is used without any specific meaning – in legal governance, transparency needs to be supported by concrete accountability measures, to create change if the conduct of target actors is not aligned with existing legal or ethical rules.

4 Overview of the Papers

The above perspectives on transparency provide a conceptual backdrop that contextualises the contributions of the papers included in this thesis. Before analysing these perspectives in light of the concept of AI transparency – which will be examined in the next chapter – the present chapter provides a brief summary of the appended papers. They are introduced in the order they were developed and submitted, each targeting journals from different disciplinary focuses.

4.1 Paper I

Explaining automated decision-making: a multinational study of the GDPR right to meaningful information

What information do companies actually reveal about their automated decisions when the right of access to information is invoked by the data subjects? I was curious to find out myself, so when the opportunity arose to empirically examine how the GDPR is implemented in practice, I readily accepted to collaborate in this research project.

Specifically, the questions pursued in the article concerned the interpretation of the right to receive 'meaningful information about the logic involved' in 'automated decision-making', as outlined in Article 15(1)(h) of the GDPR. Although this provision concerns automated decision-making in general, it is also relevant in the context of AI-driven decisions, as explained above.

The article focuses on insurance companies and their practice of applying automated decision-making to determine the rate of home insurance premiums for their customers. The research questions concerned the information that insurance companies in the EU disclose when consumers ask about the logic behind automated decisions. The study was conducted on a sample of companies from five EU countries: Denmark, Finland, the Netherlands, Poland, and Sweden.

The study began with the recruitment by the article's co-authors of volunteers-participants. The participants were later asked to send requests to their home insurance companies requesting information about how insurance premiums were set by referring specifically to Art. 15(1)(h) GDPR. The requests were sent to 26 insurers across the countries under study. I contributed by investigating the implementation of this provision within the Polish insurance market. The responses were later analysed, compared, and discussed by the co-authors.

A particularly notable finding during the course of this study was the recognition of differences in interpretation of the provision in question within the research group. This problem will be further discussed in Section 5.2. In general, however, it could only be concluded that the requirement for meaningful information about algorithmic decisions is open to varying interpretations. It was, however, unclear how the insurance companies interpreted this provision, and whether the area of the study would fall under the narrow or broad scope of interpretation.

The findings in this Paper contribute to answer RQ 1 in how individual-oriented AI transparency has been conceptualised, designed, and implemented in the EU technology regulations.

4.2 Paper II

Regulating high-reach AI: On transparency directions in the Digital Services Act

Paper II addresses the notion of *high-reach* AI, that is, AI systems which are deployed on large societal scale. These include tools like generative AI (e.g. ChatGPT) and recommender systems used by dominant platforms to rank search results, personalise news feeds, or suggest music and films. The paper highlights the significance of such systems precisely because of their widespread use: even minimal individual effects, when aggregated, may pose systemic risks comparable to those classified as high-risk under the AI Act.

This article examines a prominent example of high-reach AI: recommender systems deployed by dominant social media platforms, which represent one of the most widespread AI applications. We point to some of the negative effects which they pose on end-users and societies, and explore how the Digital Services Act (DSA) addresses them through targeted transparency provisions.

Using the theoretical framework of horizontal and vertical transparency directions by David Heald (2006), we examine transparency provisions which DSA introduces for both end-users (referred to as 'recipients of the service' in DSA terminology) and oversight bodies.[47]

On the one hand, the DSA could be seen as providing tools for increased level of monitoring of digital service practices on both the national and EU level through the introduction of regular reporting and *systemic risk assessments* by the platforms. Moreover, when necessary, the DSA grants full access to information for oversight authorities, i.e. DSCs and the Commission. Drawing from Heald's (2006) conceptual framework, we interpret this direction of information flow as *vertical upward transparency*, which appears to be significantly reinforced by the DSA.

On the other hand, the DSA framework aims to also bolster transparency measures for end-users, most notably by providing more detailed information about the profiling, recommending, advertising, and by creating the *notice-and-action* mechanism. The lack of the latter mechanism had been one of the most criticised issues in platform governance prior to DSA's adoption. Following the DSA, upon triggering of the notification procedure concerning the suspectedly illegal content, both sides of the dispute should be informed about the basis of the platform's decision, have the possibility to appeal, resort to *out-of-court* dispute resolution, as well as bring the case to the DSCs. Following Heald's conceptual framework, we conceptualise this level of transparency as *horizontal inwards transparency.*

Our analysis of the DSA transparency provisions in light of Heald's transparency directions show that the new horizontal transparency

[47] The DSA establishes a regulatory framework for digital services, entrusting primary oversight responsibilities to the Digital Services Coordinators (DSCs) at the national level, and designating the oversight of Very Large Online Platforms (VLOPs) and Very Large Search Engines (VLOSEs) to the Commission.

provisions mostly clarify or supplement the existing rights available to individuals in the context of digital services. However, the reinforcement of the vertical transparency in form of the regular systemic risk assessments and the investigative powers to the EU oversight bodies can be seen as one of the most far-reaching changes in the digital services governance framework. Nevertheless, the effectiveness of these transparency tools largely depends on the extent to which they will actually be operationalised by the oversight bodies.

The article contributes to answering of both RQ 1 and RQ 2 in the analysis of how the transparency measures concerning AI systems (recommender systems in this case) have been designed and implemented in the EU technology regulations.

4.3 Paper III

Enforcement Design Patterns in EU Law: An Analysis of the AI Act

Paper III explores how the enforcement of EU law is designed within the multinational context of the European Union. Traditionally, the enforcement of EU law has been the main responsibility of the Member States. Nowadays, however, the tasks of monitoring, investigation and sanctioning violations of EU law is increasingly centralised, often involving an active role of EU bodies, such as EU Commission, EU agencies and EU networks.

Drawing on the concept of *legal design patterns* (Koulu et al., 2021), this article scrutinises the AI Act enforcement strategy with the specific focus on transparency for oversight bodies. While the design approach to law may theoretically allow for multiple ways of identifying patterns, the analysis here centres on two overarching EU enforcement patterns, presented in Figure 1 below:

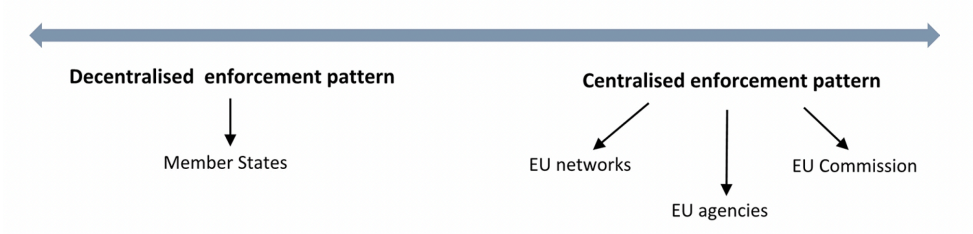

Fig. 1: Illustration of the decentralised and centralised EU enforcement patterns

The first, *decentralised enforcement pattern*, refers to the EU legal frameworks which are primarily enforced by the Member States. The second, *centralised enforcement pattern*, denotes various forms of shared and/or centralised methods of enforcement of EU laws, including enforcement by EU networks (involving the close cooperation between national authorities), EU agencies (centralised EU bodies), and by the EU Commission. In the latter model, the oversight bodies may have varying scope of competences, ranging from coordinative to fully exercising the enforcement activities, including monitoring, investigation, and sanctioning violations of the legal rules.

Against this conceptual backdrop, we analyse what enforcement patterns can be found in the AI Act enforcement structure, and what implications the choice of this enforcement model may have on the effectiveness of the AI Act.

Although the AI Office has been entrusted the task of oversight of the general purpose AI used across the Union, the Paper shows that the onus of the AI Act enforcement – including the transparency mandates – will be primarily placed on the national oversight bodies.

Our analysis points to the potential implications of the largely decentralised AI Act enforcement mechanism, which may result in uneven enforcement levels across Member States. This enforcement strategy, with primarily Member State enforcement and coordinative role of EU bodies, is reminiscent of the GDPR model, which has faced well-documented challenges. For instance, a report of the European Parliament on the assessment of the GDPR pointed to the fact that 21 DPAs explicitly stated that 'they do not have sufficient human, technical and financial resources, premises and infrastructure to effectively perform their tasks and exercise their powers' (European Parliament, 2021). In view of the potential

challenges with the AIA enforcement, the article emphasises the need for a close collaboration of the Member States on the centralised level, to prevent enforcement fragmentation exemplified by the GDPR.

The Paper provides further insights to RQ 2 by indicating that the choice of the enforcement model may have important implications on the effectiveness of EU laws.

4.4 Paper IV

High-risk AI transparency? On qualified transparency mandates for oversight bodies under the EU AI Act

Paper IV focuses specifically on the transparency measures designed for oversight bodies, which have been granted full access to relevant information to assess an AI system's compliance with the AIA. Drawing on the work of Frank Pasquale (2010, 2015) and broader AI literature, I refer to this high level of disclosure limited to a designated third party as *qualified transparency*. While the main responsibility for ensuring that the deployed AI technologies are safe and comply with the existing laws rests on the AI providers, I examine the responsibilities, obligations, and powers of the Market Surveillance Authorities (MSAs), notified bodies (NBs), and EU institutions, which under the AIA are tasked with ensuring the sound operation of the Regulation.

In the first part of the article, I explore the meaning and significance of the concept of *qualified transparency*. I then expand on this concept by drawing on broader literature on AI transparency to identify the key functions and features required for qualified transparency to be effective. In the second part of the article, I examine how this concept has been integrated into the AIA, and which national and EU institutions have been entrusted with this level of transparency in relation to AI technologies.

The analysis shows that the AI Act grants this level of transparency to the MSAs and the AI Office, and to a certain extent, to the notified bodies.

However, the role of notified bodies appears likely to remain limited in practice. This is because almost all high-risk AI systems listed in Annex III AIA may follow solely the conformity assessment procedure based on the internal control, without the involvement of NBs. The only exception are AI systems that are 'intended to be used for the 'real-time' and 'post' remote biometric identification of natural persons' (Annex III (1) AIA), yet once the harmonised standards are in place, even in those cases the providers may choose to follow the internal control procedure (Article 43(1) AIA).

Since the role of NBs is limited under the AIA – at least for now – even greater responsibility for effective implementation falls on the MSAs. The expectations placed on MSAs are therefore exceptionally high: they are tasked with ensuring that all high-risk AI systems comply with the AIA, detecting and investigating prohibited practices, and monitoring whether AI providers meet their obligations to natural persons, particularly in relation to uses such as interactive AI systems, emotion recognition, or synthetic content. Ensuring the effective enforcement of the AIA may therefore present a significant challenge for oversight authorities, in view of the complexities and the breadth of oversight responsibilities concerning all AI uses deployed on the market.

This article contributes to answering RQ 2 in analysing the conceptualisation, design, and implementation of oversight-oriented AI transparency in the EU technology regulations.

5 Analysis: Conceptual dimensions of 'AI transparency'

As stated earlier, transparency of AI systems has been seen as one of the fundamental components in fostering the Trustworthy AI governance framework in the EU (AI HLEG, 2019; European Commission, 2020b). Yet, while there seems to be a widespread agreement as to the central role of transparency in ethical AI guidelines, what exactly is meant by this principle in specific contexts is, however, less clear.

In this chapter, I present four dimensions of AI transparency that appear in the EU's AI governance discourses. I first discuss the meaning and limitations of 1) AI transparency as a *stand-alone objective* in Section 5.1. In Section 5.2., I introduce the meaning of 2) AI transparency as a *governance ideal* and 3) AI transparency as a *governance tool*. I address the research questions by examining how the concept of AI transparency has been conceptualised as a governance ideal, and how it has been designed and implemented as governance tools in relation to the EU's policymaking objective of Trustworthy AI across the GDPR, DSA and AIA. Finally, in Section 5.3 I turn to the fourth meaning of AI transparency discussed in this thesis – 4) AI transparency as a *'floating signifier'*.

Before examining the concept across the above four dimensions, it should be noted that AI transparency is still an emerging concept, and there are many possible ways of understanding it. This is partly due to the conceptual ambiguities in both terms of 'AI' and 'transparency'. On the one hand, as pointed out, there is no one agreed definition of what artificial intelligence is in the AI field (Larsson, 2021; Russell & Norvig, 2021), and on the other hand, the term 'transparency' may be understood in different ways.

Thus, due to the conceptual entanglements of both 'AI' and 'transparency', it is not clear what the term 'AI transparency' actually means. Is the meaning of 'AI transparency' adopting the conceptual complexities of both concepts, or does it mean something specific? Does it mean a theoretical concept, or transparency requirements expressed in regulations? It could also be

considered in terms of its objective, possibly meaning AI transparency as merely information disclosure, or aiming at the actual understanding of the AI systems' workings. In my view, all such conceptual perspectives can be valid and conceivable depending on what the speaker refers to. However, such conceptual ambiguities may often lead to misunderstandings when the usage of the term 'AI transparency' is not further specified.

One line of (mis)understandings surrounding 'AI transparency' could be inferred by considering the various meanings of the object to which the concept of transparency refers to – AI in this case. However, as noted above, in the literature concerning AI, the question of what exactly is meant by 'AI' is still disputed among the AI researchers. Moreover, in the governance discourses, the notion of transparency in AI is often understood as a concept encompassing various components which shape AI systems, thus not being limited to technical aspects of the systems, but also including 'human components', such as their design choices and decision-making processes. For example, Koene et al. (2019) discerns seven areas which the term 'AI transparency' may relate to. These include the transparency of data (concerning the sources of the data, how the data was pre-processed, bias verified, etc.), algorithms (relating to testing, third-party reviews), goal (clarity on competing objectives, such as safety vs. efficiency) or usage (for example, for users to know what personal data a system is using and potentially to control their data usage). In this context, Larsson & Heintz (2020) point out that it could be useful to differentiate between the terms *AI transparency* and *algorithmic transparency* for the sake of clarity in what the speaker refers to. As the authors argue, the concept of *AI transparency* (or *transparency in AI*) would take 'a system's perspective rather than focusing on the individual algorithms or components used', thus would refer to transparency of AI systems as a broader concept, including the decision-making processes, for instance. In contrast, *algorithmic transparency* would denote 'the notion of algorithms in computer science as a finite step-by-step description on how to solve a particular class of problems', which would therefore relate to transparency of a specific algorithm, or a set of algorithms, operating together within an AI system. This differentiation is important for the precision of whether the speaker refers to the idea of transparency of an AI system operating as a whole,

including the underlaying datasets, and non-technical aspects, or to the narrower notion of transparency of an algorithmic model[48].

In the broader sense, the concept of AI transparency is often referred to in regulatory and governance contexts, for instance relating to accountability frameworks. As such, AI transparency would therefore encompass 'human components' of decision-making processes. Indeed, since AI systems often embed 'human values and ideologies either inadvertently or by choice' (Koulu, 2021), the deployment of AI systems on scale may have significant societal impact on such issues as privacy, autonomy, non-discrimination, and/or the democratic discourse. Thus, a comprehensive assessment of AI systems should consider AI systems as 'algorithmic assemblages of humans and non-humans' working together (Ananny & Crawford, 2018; Kemper & Kolkman, 2019).

While keeping in mind that the object of AI transparency may relate to the broader or narrower meanings, one could also consider what is meant by the concept of *transparency* in the phrase 'AI transparency', which will be the focus of the analysis below. As mentioned, although the term 'transparency' can be understood in different ways, I focus on exploring four different ways of understanding of 'AI transparency' based on the different meanings of transparency as pointed to in Chapter 3. In the upcoming section, I present the meaning of AI transparency as a stand-alone objective.

5.1 AI transparency as a stand-alone objective

Following Koivisto's framework, one way of understanding 'transparency' – yet in relation to AI – would stem from transparency as a visual metaphor, building on the optical condition and the inherent promise of 'seeing is knowing'. In this sense, transparency could be seen as an ideal, as in Plato's theory of forms – a state of full and undistorted visibility of an object, suggesting an 'unmediated immediacy' (Koivisto, 2022). This understanding of the concept of 'AI transparency' would therefore suggest – as in the metaphorical mechanisms discussed above – that the 'truth' about AI

[48] For example, in case of algorithmic *opacity*, it might potentially be possible to address this issue by applying additional testing or xAI solutions, as mentioned above.

systems is within our reach. However, the visual metaphor of 'seeing is knowing' may promise more than it can deliver when relating to abstract or complex phenomena (Koivisto, 2022), such as AI. Thus, while the idea of fully understanding AI systems may be unrealistic, this interpretation of 'AI transparency' remains defensible in public discourse on AI — especially for audiences unfamiliar with the inherent limitations of transparency in AI technologies.

Although it may not be feasible to reach 'full AI transparency', this does not mean that such systems are entirely unknowable. It is therefore helpful to reflect on the distinction between – what I refer to as – the *knowable* and *unknowable* aspects of AI systems. The distinction between these scenarios may be useful in particular in the process of policy- and law-making, as it is important to understand which elements of AI systems could be demanded to be made explicit, tested, or disclosed, and which aspects of AI systems might go beyond what may be reasonably required of AI providers.

5.1.1 Knowable aspects of AI systems

Certain aspects of AI systems are, or could be made, knowable. Much of such information is available internally for AI developer teams and other decision-makers within AI companies. Such data could include, for instance, information on how databases were received and pre-processed, what kind of techniques were used for training of algorithms, testing methods, what the level of resulting accuracy is, which parameters the algorithms were optimised for, as well as the internal decision-making processes. As Ebers (2019) observes, to effectively debug, troubleshoot, and improve AI systems, researchers and developers need a clear understanding of how their models operate internally.

For AI companies, some of this information may be highly sensitive and is therefore typically kept confidential. In this context, Moshe Halbertal's theorising on transparency and why knowledge can be limited or impossible to access can be very useful (Halbertal, 2009; Koivisto, 2022). Halbertal distinguishes between three groups of *esoteric knowledge*[49] –

[49] According to Cambridge Dictionary, the term *esoteric* means 'very unusual and understood or liked by only a small number of people, especially those with special knowledge'. In contrast,

instrumentally, internally, and essentially esoteric knowledge – each based on different assumptions of the effects of information disclosure.

Instrumentally esoteric knowledge means a type of knowledge which is kept hidden as its disclosure could cause serious harm for the general public. In this sense, one could imagine that the information concerning AI could be concealed in the contexts of, for instance, national security or law enforcement. Roberts (2006) points out that certain 'enclaves' within governments have seen little progress in terms of transparency. He specifically refers to the security sector. As Meijer (2014) observes, it is common in many countries that specific classification systems are implemented to restrict access to sensitive information in these governmental domains. Unrestricted transparency of the systems might expose the algorithms to the risk of manipulation of the systems by users (Ananny & Crawford, 2018; de Laat, 2017). Moreover, releasing all the data to the public might affect privacy of individuals on whose data algorithmic models had been trained. AI systems often process vast amounts of personal and sensitive data, and complete transparency could expose individuals' private information, leading to breaches of privacy and potential misuse (Barocas & Selbst, 2018).

Internally esoteric knowledge, in turn, refers to the situations in which the disclosure of information would only harm the knowledge holder, as the value of the information stems from its restricted access. When applied to the area of AI, the most prominent reason for non-disclosure include the competition interests. The trade secrecy, non-disclosure agreements, and business confidentiality can be seen as building on such reasons of information secrecy. Information disclosure of the software's technical information could compromise the competitive advantage of AI companies over their competitors. AI systems often involve proprietary algorithms and technologies in which companies invested in the process of development, which would undermine incentives for innovation.

However, as pointed out earlier, in some cases the main reason for information secrecy is the intentional avoidance of oversight scrutiny (Pasquale, 2015). This strategy of concealing information about software

exoteric knowledge refers to common knowledge, accessible to the general audience (also in: Halbertal, 2009; Koivisto, 2022).

systems often operates through various legal mechanisms — what has previously been referred to as *intentional opacity* or *legal opacity* (Burrell, 2016; Söderlund, n.d.; Tschider, 2021). In such cases, the justification for secrecy may be viewed as illegitimate, as these practices aim to obscure the harmful effects or risks that AI technologies may pose.

5.1.2 Unknowable aspects of AI systems

The third category of esoteric knowledge that is relevant here as well is the *essentially* esoteric knowledge, which relates to situations when the object of knowledge cannot be meaningfully verbalised and captured (Halbertal, 2009). This may be due to the mystical nature of a given phenomenon, or because its complexity or high level of abstraction. As Halbertal puts it, in such contexts 'transparency is essentially blocked' and we can 'only intimate the truth by way of symbols and hints' (Halbertal, 2009).

It could be argued that many of the 'black-box' issues in AI stem from this perspective, as there are some aspects of AI systems that may not be known and understood about AI systems *directly* by humans, even within the AI company creating the software. Examples of unknowable aspects of AI systems could be, for instance, the inherent opacity of most advanced AI systems applying deep learning methods, or the complexity of many algorithms working together (Kemper & Kolkman, 2019). As mentioned in Paper IV, the 'black-box' AI algorithms could potentially be tested with the help of other algorithms (such as xAI). This would, however, delegate the task of 'seeing' the black-box AI to other algorithms. The question that could be raised is, therefore, whether these methods would provide a sufficient level of confidence to, effectively, mediate the task of scrutiny of black-box AI models.

Beyond the inherent opacity of many AI systems, there are additional aspects that may remain fundamentally unknowable – even for cautious developers applying all appropriate and relevant testing methods. These may include, for instance, uncertainty how an AI system might behave when exposed to different datasets[50] – an issue which could potentially introduce

[50] This point was raised, for example, in Wagner et al. (2025), who point to the difficulties AI developers face in aligning testing conditions with real-world deployment environments —

the risk of unforeseen biases or errors; what effects an AI system may have on particular individuals, groups, or society at large (despite robust impact assessments conducted), in particular in the long run; whether internal malfunctions, possibly triggered by external conditions, could cause unexpected failures (such as glitches); or whether the system's cybersecurity is sufficiently robust to withstand external threats (despite reasonable testing efforts)[51]. Still, it should be noted that the uncertainty in AI systems can also to some extent be measured, and managed, which should prompt appropriate governance responses such as risk management. Moreover, what may be regarded as 'essentially esoteric knowledge', referring to the notion of 'unknown unknowns', may actually be possible to transpose into 'known unknowns', in particular when technical possibilities of testing and quality assurance will continue to progress along with AI technologies.

In sum, it may not be a feasible objective in practice to reach the state of ideal, undistorted AI transparency, the 'whole truth' about AI systems, including full understanding of AI workings and a range of impacts in different application contexts. It may not possible for AI providers, and – most likely – even more so for oversight authorities. Still, the concept of AI transparency as a stand-alone objective could be understood as a *heuristic*. Some aspects of AI systems are known by an AI company, and some aspects may at least be specified as, for instance, a level of confidence and accuracy in different contexts.

Thus, the difference between the knowable and unknowable aspects of AI, as well as the different reasons for concealing information about the knowable aspects of AI systems are important to delineate. Both knowable and unknowable aspects of AI (i.e. the level of uncertainty that cannot be sufficiently concretised and managed), are equally important in the policy-

particularly when they lack information about where and how their systems will ultimately be used.

[51] For this reason, it seems, the AI Act states that the risk management system of high-risk AI systems should comprise of the identification and analysis of the known and the *reasonably foreseeable* risks and under conditions of *reasonably foreseeable* misuse (Art. 9 AIA). The notion of 'reasonably foreseeable misuse' is further defined in Art. 3 AIA, as 'the use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behaviour or interaction with other systems, including other AI systems.'

and decision-making processes in the broader governance contexts, which will be discussed below.

5.2 AI transparency as a governance ideal and a governance tool

The second dimension of 'AI transparency' builds on the idea of transparency as a governance ideal in law (Koivisto, 2022). In contrast to the notion of AI transparency as a stand-alone objective, AI transparency as a governance ideal would mean a regulatory approach that considers the inherent limitations of knowledge concerning technologies pointed to above, as well as takes into account important and legitimate interests of all stakeholders in the AI governance framework.

However, framing AI transparency as a governance ideal raises the question of how far transparency should extend in practice. When designing AI transparency requirements within regulatory frameworks, the extreme scenarios – of unrestricted secrecy or unrestricted transparency – are undesirable for different reasons. The first scenario envisions a situation where AI systems are completely secret, leaving no room for third-party scrutiny – not even by oversight bodies. Under such conditions, identifying harmful effects of AI, whether in the form of bias, systemic error, or discriminatory outcomes, becomes highly problematic (e.g. de Laat, 2017; Pasquale, 2010, 2015). This lack of insight also facilitates anti-competitive behaviour and weakens accountability mechanisms, as it becomes difficult to assign responsibility when harms occur (Veale & Borgesius, 2021).

The opposite point of departure – unrestricted transparency – would entail making all information about AI systems publicly accessible. In this scenario, a number of other concerns would arise as well (e.g. de Laat, 2017; Pasquale, 2010, 2015). As touched upon earlier in the context of why certain knowledge remains instrumentally or internally esoteric, this scenario would be harmful for many legitimate interests in the governance frameworks, such as safeguarding competition incentives, protecting systems from user manipulation, and avoiding significant privacy and security risks.

In light of the above competing interests, it is necessary to recognise that both complete secrecy and unrestricted transparency of AI systems entail undesired consequences for governance frameworks. As Turilli & Floridi (2009) observe, both scenarios are extreme approaches which would fail to foster positive ethical implications while risking the promotion of negative ones.

The ideal of transparency in AI governance frameworks would therefore need to balance the conflicting interests by adopting a nuanced approach to transparency, taking into account the specific context (Kaminski, 2020; Pasquale, 2015). The information to be disclosed must be carefully considered by evaluating its potential consequences for other stakeholders.

In what follows, I explore how AI transparency has been envisioned as a governance ideal that supports the goal of Trustworthy AI. This involves analysing how transparency has been conceptualised and designed for two stakeholder groups — individuals and oversight bodies — within the GDPR, the DSA, and the AIA. The sections also consider how these governance tools are being implemented in practice.

5.2.1 How has individual-oriented AI transparency been conceptualised, designed, and implemented in the GDPR, DSA and AIA?

a) The concept of individual-oriented AI transparency

In general, the EU policy documents highlight the importance of transparency and control over the way personal data is processed in the digital environment. When AI technologies are involved, these documents stress that the public must be informed when AI is being — or has been — used, and should have a basic understanding of how these systems operate.

Back in 2012, for example, regarding the transparency of personal data processing, the Commission emphasised the aim of improving 'individuals' ability to control their data' (European Commission, 2012b). The idea of individuals being in charge of their data in the 'new digital environment' was envisioned where 'individuals have the right to enjoy effective control over

their personal information'. The Commission also provided a strong statement for the need of 'reinforcing the right to information so that individuals *fully understand* how their personal data is handled' (European Commission, 2012b) (emphasis added), and further, 'to enable them to exercise their rights more effectively'. The responsibility for ensuring that individuals may exercise their rights was placed on oversight bodies, which are expected to be 'properly equipped to deal effectively with complaints'. Moreover, as the Communication stated, appropriate remedies should be available to individuals when their rights are violated (European Commission, 2012b).

In the context of digital services framework reform, the EU policymakers highlighted the need for improving users' safety online and protection of their fundamental rights. The idea behind the DSA was to address 'the particular impact of very large online platforms on our economy and society', and to set 'a higher standard of transparency and accountability on how the providers of such platforms moderate content, on advertising and on algorithmic processes' (European Commission, 2020a). The envisaged policy measures were meant to enhance 'user agency in the online environment', as well as the exercise of other fundamental rights such as the right to an effective remedy, non-discrimination, and the protection of personal data and privacy online. The amendments in the framework were also expected to 'mitigate risks of erroneous or unjustified blocking speech' and 'stimulate the freedom to receive information and hold opinions' (European Commission, 2020a).

Apart from underlining the role of user agency, safety, control and 'full understanding' of how personal data is processed, EU policy documents consistently emphasised the need for *trust* 'in the online environment'[52], 'in

[52] The Explanatory Memorandum to the GDPR states that '[b]uilding trust in the online environment is key to economic development. Lack of trust makes consumers hesitate to buy online and adopt new services (European Commission, 2012a).

the digital services'[53], or 'in the digital economy'[54]. This pattern appears consistently across the EU policy documents. For instance, in the Agenda for Digital Europe we read that 'Europeans will not embrace technology they do not trust - the digital age is neither 'big brother' nor 'cyber wild west''(European Commission, 2010). The reference to the value of trust in the legislative contexts is, in itself, a noteworthy trend — one that has become increasingly observable in EU legal development over the past three decades (Chamberlain & Kotsios, 2025).

With regard to AI technologies specifically, the EU Commission emphasised that 'EU citizens have the right to know that AI systems they are using are not affecting them in negative ways, that they are not discriminatory or erroneous, to know what data is processed, and that no other data is processed about them' (European Commission, 2018b). The objective of this level of transparency has been projected as 'to enable individuals to understand the way algorithms work in general terms' and – again – to 'strengthen trust' (European Commission, 2018a) of the public in AI technologies. Moreover, the Commission's White Paper on AI points to the importance of the public being informed whenever they interact with an AI software, except in cases where such interaction is 'immediately obvious' to citizens (European Commission, 2020b).

Thus, the recurring themes in the EU policy documents discussing the AI governance objectives for individuals would emphasise the importance of agency, control, information, safety, and trust. However, the policy documents would not reveal more closely how these values should be expressed in the legislative frameworks.

Academic literature on the topic highlights the importance of this level of transparency in promoting AI literacy among the general public as well. For example, Burrell (2016) observes that algorithmic design and coding are highly specialised skills, largely inaccessible to most people. Thus, as

[53] In the Digital Single Market Strategy for Europe, the Commission states that '[t]he General Data Protection Regulation will increase trust in digital services, as it should protect individuals with respect to processing of personal data by all companies that offer their services on the European market.' (European Commission, 2015).

[54] The European Parliament's resolution on improving the functioning of the Single Market from 2019 calls for 'measures which ensure consumer trust in the digital economy' (European Parliament, 2019b).

Edwards & Veale (2017) point out, AI transparency for individuals should support the development of more accurate 'mental maps' of how AI models function, thereby serving a pedagogical purpose. In addition to enhancing understanding (IEEE, 2022), transparency should enable individuals to exercise their rights — such as the rights to a fair trial, non-discrimination, and privacy. Edwards and Veale (2017) further stress that individual-oriented transparency measures are essential to fostering public trust and encouraging the meaningful use of machine learning systems.

Apart from the objectives of transparency for individuals, both EU policy documents and the literature frequently address *how* information about AI systems should be communicated to individuals. The EU Commission's *Shaping Europe's digital future* emphasises that information concerning AI systems should be 'objective, concise and easily understandable', and the way in which the information is to be provided 'should be tailored to the particular context' (European Commission, 2020b). In particular, with regard to Art. 22 GDPR-decisions, Kaminski (2020) observes that the individual transparency as an explanation of a decision should be provided 'at an abstract enough and simple enough level so as to be understandable, but also complex enough to be actionable, to allow her to contest the decision.'

What emerges from the above analysis is that EU policy documents conceptualise AI transparency for individuals in terms of both access to information and control. Individuals are not only to be informed about what data is processed and for what purposes, but are also expected to exercise 'full control' over their data. In the context of AI systems, this includes knowing when one is interacting with (or is subject to) an AI system, and having a general understanding of how the systems work and what their limitations are — elements deemed crucial for fostering Trustworthy AI environment.

The limits of individual-oriented AI transparency

The concept of AI transparency, even when framed as a governance ideal, is not without limitations. These limitations can be broadly grouped into three categories. The first category, frequently emphasised in the EU policy documents and the scholarly literature, arises from the need to balance transparency with competing rights and interests. For instance, as the EU

policy paper *Safeguarding Privacy in a Connected World* (European Commission, 2012b) states, the right to the personal data protection is not an absolute right, as it 'must be considered in relation to its function in society and be balanced with other fundamental rights, in accordance with the principle of proportionality'[55]. Most commonly, the individual-oriented transparency restrictions are justified by the trade secrecy and business confidentiality (Burrell, 2016; Foss-Solbrekk, 2021; Pasquale, 2015; Tschider, 2021).

The second group of constraints of individual-oriented transparency can be linked to the limits of human cognitive capacity (see, for instance, Zech, 2023). Similarly, Edwards & Veale (2017) point out that individuals are simply 'mostly too time-poor, resource-poor, and lacking in the necessary expertise to meaningfully make use of these individual rights'.

The third type of limitations relates to the broader question of what transparency for individuals can realistically achieve within the governance framework. One of such limitations relate to the argument that individual-oriented transparency alone cannot solve the issue of accountability (Ananny & Crawford, 2018). Placing the burden on individuals to, interpret and evaluate information does not constitute an effective governance tool, in particular since end-users are disconnected from power (Edwards & Veale, 2017). As Kaminski (2020) observes, individualized explanations of AI decisions 'don't empower people, and instead distract from more effective ways of governing'. Ananny & Crawford (2018) describe this overreliance on individual rights as a 'neoliberal models of agency', and Edwards and Veale (2017) as a 'transparency fallacy' — the illusion that individual rights alone enable meaningful control over machine learning systems. Even if transparency for individuals is meant to contribute to the accountability mechanisms as in Meijer's *indirect route: strengthening vertical accountability,* such measures remain ineffective without a functioning forum for redress to (cf. the notion of the 'critical audience' in Kemper & Kolkman, 2019).

[55] In line with Article 52(1) of the Charter, limitations may be imposed on the exercise of the right to data protection as long as the limitations are provided for by law, respect the essence of the right and freedoms and, subject to the principle of proportionality, are necessary and genuinely meet objectives of general interest recognised by the European Union or the need to protect the rights and freedoms of others.

Moreover, as pointed out above, the individual-oriented transparency is insufficient to address the systemic risks of algorithmic systems deployed at scale. Rights granted to individuals — such as those under the GDPR, DSA, and AI Act — are not 'scalable' themselves. That is, they may fail to capture the aggregated or collective impacts of *high-reach AI systems* (Söderlund et al., 2024) – as discussed in Paper II – such as social media recommender systems, as these are not identifiable from the individual's perspective. Instead, as Crawford & Schultz (2014) highlight, algorithmic harms typically arise from how systems classify groups, which may also mean that the way algorithmic classification operates may prevent certain individuals from realising the opportunities they might have otherwise had.

In sum, while the scholarly literature highlights the importance of individual-oriented transparency in AI, it also raises concerns about the overreliance on individual rights to fulfil broader objectives such as accountability, non-discrimination, and public safety (Ananny & Crawford, 2018; Busuioc et al., 2023; Edwards & Veale, 2017). In response to these concerns, there has been an increasing demand for a comprehensive governance framework for AI — one that can both prevent unlawful data use and harmful AI impacts, and protect individual's rights.

b) The design of individual-oriented AI transparency

The preceding discussion on the objectives of AI transparency for individuals is reflected in the way transparency has been articulated across the three regulatory frameworks analysed in this thesis. In the following, I provide a short overview of the relevant provisions in the GDPR, the DSA, and the AIA that aim to strengthen AI transparency for individuals.

General Data Protection Regulation (GDPR)

In general, under the GDPR, transparency features as one of the overarching personal data processing principles. According to Article 5 GDPR, personal data should be 'processed lawfully, fairly and in a transparent manner in relation to the data subject'.

Other provisions relevant for individuals (that is, data subjects under the GDPR), include various rights that data subjects may invoke vis-à-vis the

data controllers[56]. On the basis of Articles 12-14 GDPR, data controllers are required to proactively provide specific information to data subjects 'in a concise, transparent, intelligible and easily accessible form, using clear and plain language'(Art. 12 GDPR). Organisations processing personal data must provide such information as the identity and contact details of the controller, data protection officer, purposes and legal basis for processing, recipients of data, and any international transfers. Additional details include data retention periods, individuals' rights (access, rectification, erasure, objection, portability), complaint procedures, and – importantly – details on automated decision-making or profiling.

Moreover, following Article 15 GDPR, data subjects have a right to access the processed information on request. This includes information about the 'existence of automated decision-making', about the 'meaningful information about the logic involved' and 'the significance and the envisaged consequences' of such processing.

Article 22 GDPR, furthermore, in principle forbids fully automated individual decision-making, that is, when the processing of personal data takes place without (substantial) human involvement, and in contexts when data processing may have a significant impact on a person's life. However, such decisions may be still lawful if they are necessary for entering into or performance of a contract, when such decisions are authorised by EU or national laws, or on the basis of an *explicit* consent provided by the data subject. In such cases, the data controller should implement suitable measures to safeguard the data subject's rights, freedoms and legitimate interests. These should include, as the passage states, 'at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision' (Art. 22(3)(c) GDPR). Notably, in this context, Recital 71 GDPR also mentions a right 'to obtain an explanation of the decision reached after such assessment' as part of suitable measures to safeguard the data subject. However, since recitals mainly function as guidelines on how to interpret law but are not strictly binding themselves, the existence and the content of a 'right to (an)

56 Under the GDPR, data controllers are natural or legal persons which determine the purposes and means of the processing of personal data (Art. 4 GDPR).

explanation' is still disputed in the legal scholarship (see, for instance, Edwards & Veale, 2017; Kaminski, 2021; Wachter et al., 2017).

In addition to the above doctrinal debates, as mentioned, the study in Paper I has revealed an ambiguity in the different language versions of the GDPR. The issue concerns the reading of Art. 15(1)(h) GDPR, and a sample of the different language versions of the GDPR – exemplified by the English and Swedish version – is presented in Figure 2 below:

Article 15	Artikel 15
1. The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:	1. Den registrerade ska ha rätt att av den personuppgiftsansvarige få bekräftelse på huruvida personuppgifter som rör honom eller henne håller på att behandlas och I så fall få tillgång till personuppgifterna och följande information:
h) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, *at least in those cases*, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.	h) Förekomsten av automatiserat beslutsfattande, inbegripet profilering enligt artikel 22.1 och 22.4, *varvid det åtminstone I dessa fall* ska lämnas meningsfull information om logiken bakom samt betydelsen och de förutsedda följderna av sådan behandling för den registrerade.

Fig. 2: Article 15(1)(h) of the GDPR in both English and Swedish language versions (emphasis added).

In essence, the differences in reading of the above provision may imply that the right to the 'meaningful information…' is mandatory in all cases of automated decision-making or only in cases of fully automated decisions with significant effects on individuals (by referring to the scope of Art. 22 GDPR). Specifically, the narrow reading of Article 15(1)(h) interprets 'at least in those cases' as referring to the full phrase 'automated decision-making, including profiling, referred to in Article 22(1) and (4)'. This interpretation would impose a duty on data controllers to provide the 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject' only when

data processing clearly falls within the scope of Article 22 GDPR. The broad reading would extend this obligation to all automated decision-making and profiling, not just those captured by the scope of Article 22 GDPR.

Digital Services Act (DSA)

As discussed in Paper II, on the basis of the DSA, users of digital services have been granted additional transparency rights, which were not available to them under the E-Commerce Directive. Amongst the most notable DSA provisions in this regard, the service providers should now include in their terms and conditions information 'on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making' (Art. 14 (1) DSA). Such information should, moreover, be provided 'in clear, plain, intelligible, user-friendly and unambiguous language' (Art. 14 (1) DSA).

With regard to the recommender systems specifically, the DSA obliges very large online platforms (VLOPs) to disclose the *main parameters* used in their recommender systems and explain why the specific information is suggested to the user. Following Article 27 DSA, this includes the 'criteria which are most significant' in determining the presented content, and 'reasons for the relative importance of those parameters'.

Moreover, the DSA strengthens user involvement in the mechanism of tackling illegal content. When platforms restrict certain content, users impacted by such actions must receive a justification, and have access to internal complaints systems and independent dispute resolution (Arts. 17-21 DSA).

In addition, online platforms displaying advertising must also ensure that the recipients of the service receive meaningful information on the main features used in displaying advertisements (Art. 26 DSA).

Artificial Intelligence Act (AIA)

Finally, the AIA establishes a few additional transparency measures in contexts when AI technologies are used. On the most general level, Article 50 AIA requires that individuals be informed when they interact or are exposed to AI systems. That includes such AI applications as chatbots, AI-generated content, emotion recognition systems and biometric categorisation systems.

Moreover, Article 86 AIA specifies that an individual should have the 'right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken', in cases of decisions taken on the basis of the output from a high-risk AI system listed in Annex III AIA, and which produce legal effects or similarly significantly affect individuals. This provision echoes Article 22 GDPR, and its adoption may have been influenced by the unsuccessful implementation of the right to explanation under the GDPR.

It should also be noted that the AI Act includes additional transparency provisions that may also be understood as forms of 'transparency for individuals' — in particular, measures related to instructions for use and human oversight aimed at professional users of AI systems. However, as previously stated, this category of stakeholders falls outside the scope of analysis in this thesis.

The different depths of AI transparency for individuals

Across the three legal frameworks discussed above, transparency rights can also be analysed from the perspective of the individual in relation to the AI provider. When considered together, the transparency provisions under the GDPR, DSA, and AIA can be grouped into at least three sub-categories. This division is not clear-cut, but it helps illustrate how transparency obligations shift depending on context.

Seen from this perspective, the first group would concern the disclosure of information to the general public, when no formal connection between the specific individuals and AI providers is established. While in most cases information disclosure is not legally required, AI providers may still choose to communicate certain information on their products or services proactively, for example for branding or marketing purposes. However, the notification obligations in Art. 50 AIA may also fall into this category. For instance, disclosures about the use of synthetic content, chatbots, or emotion recognition technologies in public spaces are addressed to the general public rather than to specific individuals.

The second group of transparency provisions concern situations when a formal relationship between the AI providers and individuals (e.g. as data subjects, recipients of service, or natural persons under the AIA) is established. This relationship could be based on contractual agreements

(such as concerning the use of a digital service), consent, legitimate interest, or other legal bases recognised by the relevant laws (see, for example, Art. 6 GDPR on the lawful legal bases for personal data processing). When such a formal relationship exists, individuals are entitled to a higher level of AI transparency. On the basis of the GDPR, DSA and AIA, this level of transparency entitles individuals to receive or request specific information, as outlined above.

The third category concerns situations when individuals are entitled to receive *meaningful information* about an automated decision. This level of transparency is provided to individuals who may be legally or otherwise significantly affected by fully automated decisions. This level of individual-oriented transparency could therefore be seen in Art. 22 GDPR and Art. 86 AIA. A related right to explanation — albeit in a different regulatory context — can also be found in the DSA, in cases involving content or account restrictions mentioned above (Arts. 17-21 DSA).

c) Implementation of individual-oriented AI transparency

So far, I have analysed how AI transparency for individuals has been envisioned in EU policymaking documents and how these visions have been translated into legal provisions across the GDPR, DSA, and AIA as governance tools. In this section, I turn to some aspects of their implementation.

For the first group of individuals above – that is, individuals without any formal relationship with the AI provider – transparency of AI systems is provided on the most general level. This typically involves making certain information publicly available, such as what data is collected from end-users and how the AI system operates. For example, a film streaming platform might publish an overview of its data practices before any contractual relationship is established. This might specify that it collects information such as watched content, device used, and time of access, while explicitly noting that it does not collect certain personal attributes like age or gender.

However, it can generally be observed that the higher the level of transparency designed for individuals, the more resisted its implementation

appears to be in practice. As pointed to above, the empirical findings presented in Paper I show that insurance companies do not implement the obligation to provide 'meaningful information about the logic involved' in automated decision-making in a meaningful way. The study found that the responses from companies were typically vague and incomplete. The cross-country comparison of the identified differences did not seem to follow any obvious pattern, such as the nationality, size, or ownership type. However, Danish companies tended to disclose more data points and to be clearer about the rationale of the data processing, while Polish and Dutch companies often cited business confidentiality to justify limited explanations. Moreover, a few companies failed to comply with the GDPR by not providing any response within the specified period or by not responding at all.

It should be observed, however, that the subject matter of Paper I did not clearly fall within the scope of Article 22 GDPR, thus the insurance companies were not necessarily obliged to disclose 'meaningful information about the logic involved' in their automated decision-making. At the time of the study, there was no case law from the Court of Justice of the European Union (CJEU) interpreting Art. 15(1)(h) GDPR. However, in February 2025, the CJEU issued its first judgment addressing this provision in *CK v Dun & Bradstreet Austria* (C-203/22)(CJEU, 2025), confirming the existence of a right to explanation under the GDPR:

> [I]t is apparent from recital 71 of the GDPR that, where the data subject is the subject of a decision which is based solely on automated processing and which significantly affects him or her, that *data subject must have the right to obtain an explanation* of that decision. As the Advocate General observed in point 67 of his Opinion, it must therefore be held that Article 15(1)(h) of the GDPR *affords the data subject a genuine right to an explanation as to the functioning of the mechanism involved in automated decision-making* of which that person was the subject and of the result of that decision (para. 57) (emphases added).

The CJEU clarified that under Art. 15(1)(h) GDPR — at least in cases involving credit profiling — the requirement to provide 'meaningful information' refers to disclosing the actual procedure and principles applied in the automated decision-making process. The judgment also addressed the extent to which trade secrecy may justify withholding such information. While data controllers may invoke such rights, they are now obliged to disclose the allegedly protected information to the relevant oversight body

or court, which must then balance the competing rights and interests. Although the Court held that this balancing act should be conducted on a case-by-case basis (para. 75), the ruling still sets an important precedent for interpreting Art. 15 GDPR in the contexts of credit profiling.[57]

While this is the first CJEU ruling concerning the interpretation of Art. 15 (1)(h) GDPR, the Court has already clarified that credit profiling decisions fall within the scope of Art. 22 GDPR. In the 2023 *SCHUFA Holding* judgment (C-634/21), the CJEU held that an automated credit scoring process constitutes an 'automated individual decision' due to its decisive influence on the outcome. In this case, the Court rejected the argument that trade secrets could justify withholding such information, emphasising the importance of transparency and individual rights under the GDPR.

With regard to the implementation of the DSA, the study conducted in Paper II concerning recommender systems used by VLOPs could not be subject to a similar 'reality-check'. However, after Paper II was published, the EU Commission has started to investigate the way VLOPs' recommender systems suggest the content to end-users (as per Art. 27 DSA). For example, the Commission has requested Amazon to provide detailed information on its compliance with the provisions concerning transparency of the recommender systems, including 'the input factors, features, signals, information and metadata applied for such systems and options offered to users to opt out of being profiled for the recommender systems'(European Commission, 2024b). The company has also been requested to provide additional information on the design, development, deployment, testing and maintenance of the online interface of Amazon Store's Ad Library, together with supporting documents regarding its risk assessment report (European Commission, 2024b).

In summary, the implementation process of the GDPR and DSA shows tension between the transparency rights of individuals and the interests of

[57] The above precedent may also prove significant in relation to a complaint filed against a bank in Sweden. The privacy advocacy organisation Noyb has submitted a complaint to the Swedish Data Protection Authority (IMY) concerning Swedbank. The case involves the bank's rejection of a Swedish citizen's access request on the grounds that the logic behind its credit rate constitutes a 'trade secret' (nyob, 2025b).

companies. Still, as pointed out earlier, even as a governance ideal, AI transparency for individuals faces limitations, such as trade secrecy or business confidentiality. Moreover, individuals are unable to verify whether the information provided by companies is accurate or complete. Although companies can and should continue to improve transparency practices, individuals are primarily concerned with the fairness of automated decisions and the ability to understand and challenge them.

5.2.2 How has oversight-oriented AI transparency been conceptualised, designed, and implemented in the GDPR, DSA and AIA?

a) The concept of oversight-oriented AI transparency

With regard to AI transparency for oversight purposes, EU policy documents and academic literature have consistently highlighted its fundamental role across the data protection, digital services, and AI regulation contexts.

In the data protection area, for example, the Commission emphasised the need to enhance the independence and powers of national data protection authorities (DPAs) 'to enable them to carry out investigations', and called on Member States to 'provide them with sufficient resources to do so' (European Commission, 2012b). In the digital services domain, the EU Commission stated that the single market for digital services requires cooperation between the Member States to 'guarantee effective oversight and enforcement' (European Commission, 2020a). Moreover, the DSA Explanatory Memorandum adds that the oversight tasks should also involve 'close monitoring' by the EU and national authorities of the transparency requirements imposed on the providers 'to make sure the information requirements are respected' (European Commission, 2010).

With regard to enforcement of rules specifically concerning AI technologies, the EU Commission emphasised the importance of effective application and enforcement of existing EU and national legislation (European Commission, 2020b). In the White Paper, the Commission presents this level of transparency as follows:

> In order to ensure that AI is trustworthy, secure and in respect of European values and rules, the applicable legal requirements need to be complied with in practice and be effectively enforced both by competent national and European authorities and by affected parties. Competent authorities should be in a position to investigate individual cases, but also to assess the impact on society.

The Commission further recognises the problem that the key characteristics of AI create challenges for ensuring the proper enforcement of EU and national legislation. As the Commission states, '[t]he lack of transparency (opaqueness of AI) makes it difficult to identify and prove possible breaches of laws, including legal provisions that protect fundamental rights (…)'(European Commission, 2020c). Furthermore, the Commission states that competent authorities should have the possibility 'to test such applications', in particular in cases where risks to fundamental rights are involved.

Similarly, the academic literature has consistently highlighted the need for more rigorous oversight of AI technologies. As discussed in Paper IV, AI transparency advocates have long called for the possibility of institutional intervention in cases when intentional opacity created by corporations would preclude public scrutiny. For this role to be effectively conducted, such investigation should not be restricted by the technical or legal opacity of the systems. As Frank Pasquale phrases this, '[a]gencies ought to be able to "look under the hood" of highly advanced technologies' (Pasquale, 2010). In Pasquale's work (2010, 2015), this kind of information disclosure to third parties under a duty of confidentiality is described — as noted earlier — as *qualified transparency,* since this level of transparency 'should be qualified in order to protect important intellectual property interests' (Pasquale, 2010). As Pasquale (2015) argues, such meaningful investigations require entrusting a small group of experts with full access to relevant information – unrestricted by technical or legal barriers, but bound by confidentiality rules. Kaminski (2020) refers to this approach as *systemic transparency*, which aims to reveal patterns of error, bias, and discrimination in both human and machine systems, so they can be duly addressed and mitigated.

This level of AI transparency carries the important objective of protecting individuals against the harmful effects of AI technologies. Unlike other stakeholder groups, oversight authorities are equipped with state-backed enforcement powers, positioning them as the only stakeholders capable of imposing sanctions and ensure compliance, making their role crucial in

the regulation of AI. As Knowles & Richards (2021) put it, there is a need for a 'sanction of authority' that may 'weed out' AI systems which are harmful. The authors emphasise that the public trust in AI systems cannot be forged by individuals 'through careful and ongoing assessment of their trustworthiness', but through trusting that the democratic public governance has appropriate mechanisms for ensuring trustworthiness of AI systems. In this context, oversight bodies serve as institutional guardians of the legal framework, which is expected to enable organisations to develop and deploy trustworthy AI systems. Conversely, when AI systems prove untrustworthy in practice, it signals a failure of institutional mechanisms to adequately regulate AI (Knowles & Richards, 2021).

With regard to the issue of legal opacity pointed to above – there are valid reasons for keeping the inner workings of AI systems hidden from the outside view, yet such constraints should not apply in the context of oversight investigations. While the public disclosure of such information would entail significant risks, this argument seems less convincing when applied in relation to oversight authorities (see e.g. de Laat, 2018).

Moreover – and in line with the limits of the individual-oriented transparency discussed above – it is the regulators that are in the position to monitor the broader societal impacts of AI system deployment and use at large scale. As argued in Paper II, high-reach AI systems such as of recommender systems may inflict a range of negative effects, such as to amplify the spread of misinformation, fake news, to sway political elections and polarise societies. On the individual level, such sophisticated algorithms may be very privacy-intrusive, affect user's autonomy and have a negative long-term impact on their mental health. As stated earlier, the possibility to grasp the 'broader picture' of such negative impacts of AI technologies on the societal level is particularly relevant in case of AI technologies, as due to their wide societal use it is difficult from the individual level to monitor the aggregate effects of AI systems.

The limits of oversight-oriented AI transparency

Despite the broad conceptual framing of AI transparency for oversight bodies, it has also significant limitations even as a governance ideal. As discussed in Paper IV, the issue of the technical opacity of AI – or the 'black-box problem' – still remains, even if the oversight bodies are sufficiently

funded and have at their disposal sufficient technical and socio-legal expertise. While the AIA contains certain transparency requirements vis-à-vis the high-risk AI and GPAI, the low-risk AI systems (that is, the majority of AI systems) are not subject to any minimum transparency requirements.

The issue of technical opacity and complexity of the most advanced AI, and the distinction between the *knowable* and *unknowable* aspects of AI systems has been pointed to in Section 5.1. above. Some of the oversight challenges may still arise from the technical complexity inherent in the most advanced AI systems.

In sum, the governance ideal of AI transparency for oversight bodies is commonly framed in EU policy documents and academic literature as essential for the effective enforcement of existing laws. It is seen as a safeguard against allowing technical or legal opacity to become a justification for non-disclosure (and non-compliance). In high-stakes settings, in particular, oversight bodies carry the responsibility of protecting the safety and fundamental rights of EU citizens. In such cases, AI systems should be demonstrably safe and fully compliant with relevant regulations. However, in low-risk AI systems, where accuracy is not compromised by transparency requirements, it may be more difficult to comprehensively test AI systems in practice.

b) The design of oversight-oriented AI transparency

How the oversight framework for AI systems could be designed under the EU multinational legal system has been the main topic of Paper III, which discussed the various *enforcement design patterns* of EU laws. As mentioned in Section 4.3, the analysis examined how EU laws may be enforced either predominantly at the national level or with stronger involvement of the EU institutions. The European Commission and the European Parliament generally favour centralised enforcement, as it allows for more effective oversight and accountability. However, the extent to which enforcement is centralised often depends on the position of the Member States, who typically seek to preserve their own enforcement powers. Ultimately, the degree of centralisation is shaped by the level of

political acceptance among Member States to delegate enforcement powers to EU bodies.

While the discussion in Paper III focuses on the way the enforcement tasks of EU laws can be allocated between Member States and the EU institutions, in what follows, I examine how AI transparency has been expressed in the GDPR, DSA and AIA as governance tools. As in the context of governance tools for individuals, I reshuffle the various transparency measures designed for oversight bodies, which have been provided to oversight bodies in the GDPR, DSA and AIA. This time, however, I will sort the transparency mandates following two modes of oversight procedures. The first one – which I call *standard oversight procedure*s – covers the regular mode of oversight processes. The second one – which I refer to as *investigative oversight procedure* – relates to the operational mode adopted when oversight bodies initiate formal investigations.

Standard oversight procedures

In this mode of oversight, oversight bodies operate following standard procedures of monitoring legal compliance. Such procedures involve, for instance, receiving from AI providers information such as reports or other documentation that is required under the relevant legal frameworks. This procedure could be linked to the idea of *controlled transparency* by Koivisto (2022).

Under the GDPR, data controllers are not subject to routine or periodic reporting obligations to Data Protection Authorities (DPAs), but they are required to report under specific circumstances. One example of such a circumstance is the obligation to notify personal data breaches (Art. 33 GDPR). Another obligation to report to the DPAs arises under Article 36, which requires controllers to consult the DPA prior to processing if a Data Protection Impact Assessment (DPIA) indicates a high risk that cannot be mitigated. Moreover, controllers are required to maintain records of processing activities under Article 30, and although these records do not need to be submitted proactively, they must be made available to the DPA upon request.

As further analysed in Paper II, under the DSA, intermediary service providers have specific reporting obligations designed to enhance transparency and accountability in the digital environment. All intermediary

providers, including hosting services and online platforms, must publish annual transparency reports detailing content moderation practices, including information on 'any use made of automated means for the purpose of content moderation' (Art. 15(1)(e) DSA), and actions taken based on terms and conditions (Art. 15 DSA). On top of these, VLOPs and VLOSEs are subject to additional obligations, including conducting and reporting on systemic risk assessments (Art. 34 DSA), submitting annual independent audits (Art. 37 DSA), and providing access to data for vetted researchers and the Commission (Art. 40–41 DSA). They must also maintain repositories of advertisements and report on recommender system functionality and user options.

Finally, the AI Act also introduces for providers of high-risk AI systems several reporting obligations to national market surveillance authorities (MSAs) and the EU Commission. For instance, providers must maintain comprehensive technical documentation and post-market monitoring plans (Arts. 11 and 61 AIA), which must be made available to MSAs upon request. Serious incidents or malfunctioning that could constitute a breach of fundamental rights must be reported to the competent authority without undue delay (Art. 62 AIA). Furthermore, AI providers must keep logs of the system's operation to support post-market investigation and auditing (Art. 12 AIA). The Commission, in turn, maintains a public EU database of high-risk AI systems (Art. 60 AIA), and may request information where harmonised application of the regulation is at stake (Art. 63 AIA). As discussed in Paper IV, some categories of high-risk AI systems require a pre-marketing assessment carried out by notified bodies (NBs). Yet, it is noteworthy that unlike MSAs, notified bodies are not provided the possibility to access the source code, suggesting a more limited level of transparency in their oversight role.

Investigative oversight procedures

This mode of oversight activity, as the name suggests, is at play when investigations are launched in response to suspected breaches of the law. In Koivisto's (2022) terminology, it can be understood as a manifestation of *uncontrolled transparency* tools.

Such transparency tools can be identified across the GDPR, DSA, and AIA, and these frameworks provide oversight authorities with considerable

investigative powers. In the GDPR, under Article 58, the national supervisory authorities have been granted investigative, corrective, and advisory powers. Among other transparency mandates, they can require data controllers to provide necessary information, conduct data protection audits, and access any premises. Administrative fines may also be imposed, either alone or in combination with other corrective actions, depending on the nature and severity of the violation.

In the DSA, an equivalent provision is laid down in Article 51, which states that the national authorities – Digital Services Coordinators (DSCs) – have the power to require to provide any information, to carry out inspections of any premises in order to examine, seize, take or obtain copies of information relating to a suspected infringement in any form, and the power to ask any member of staff to give explanations. With regard to the VLOPs and VLOSEs, such investigation powers are vested in the EU Commission. If the Commission concludes that a VLOP or VLOSE has breached the DSA, it can impose fines of up to 6% of the company's global turnover.

With regard to the powers of oversight bodies under the AIA, as discussed in detail in Paper IV, they have been granted far-reaching transparency mandates as well. On the national level, the scope of purview of the national market surveillance authorities (MSAs) covers all AI providers, regardless of the level of risk the AI systems pose. Notably, under the market surveillance regulation (Regulation 2019/1020), the investigation powers of the MSAs to realise their responsibilities are very broad, as these may include the authority to 'require economic operators to provide relevant documents, technical specifications, data or information on compliance and technical aspects of the product', the power to carry out unannounced on-site inspections, and enter any premises in the course of investigations. With regard to the high-risk AI systems, moreover, the MSAs have been granted full access to the documentation and datasets used for the training, validation and testing of high-risk AI systems, as well as access to the source code upon a reasoned request (Art. 74 AIA). Similar enforcement powers have been provided to the AI Office when acting in its role as the market surveillance authority concerning the GPAI, as pointed to in Paper IV.

What is important to highlight here is that the AIA also provides additional avenues for national authorities supervising the obligations to respect

fundamental rights protection laws – including the oversight of the data protection laws by DPAs. Such authorities have been granted under the AIA similar transparency powers with regard to the high-risk AI systems listed in Annex III AIA, including documentation created for the AIA purposes, and the possibility to request the MSAs to organise additional testing of the AI system when the information provided is insufficient to determine whether an infringement concerning fundamental rights has occurred.

In sum, the analysis of the powers of oversight authorities under the GDPR, DSA, and AIA, in particular in the context of investigations, shows that these are indeed far-reaching, thus this could be seen as the highest level of AI transparency available to third-parties in the AI governance framework. As pointed out in Paper IV, the investigations may also concern information that is normally protected by such legal measures as trade secrecy, to verify whether the claims about AI systems are true. Moreover, oversight bodies have the power of sanctioning and enforcement in bringing the unlawful practices of AI providers to account. However, as noted in Paper IV, while the transparency mandates granted to oversight bodies provide the possibility to scrutinise AI systems, their effectiveness ultimately depends on the discretion of the authorities whether to intervene. Thus, even *qualified transparency* granted for the oversight purposes does not automatically lead to accountability measures. What may ultimately make or break the effectiveness of these rules is the extent and the way in which the above transparency mandates are implemented by the oversight bodies.

c) Implementation of oversight-oriented AI transparency

Having explored how EU policy documents and the scholarly literature conceptualise AI transparency for oversight bodies, and how these ideas have been subsequently designed into binding rules across the GDPR, DSA, and AIA, this section turns to questions of their implementation. However, as pointed out above, the analysis will not be exhaustive. I will point to several issues discussed in Paper III concerning the implementation of the GDPR, which may shed light on the possible challenges with the AI Act enforcement.

As the GDPR has been in force for many years now, Paper III discusses the implementation of the GDPR enforcement mandates by the oversight bodies. As already stated, the shortcomings in the GDPR implementation have been acknowledged by the EU and national institutions themselves. Despite the GDPR's aim to enhance data protection (as stressed in the EU Commission's Communication 2012b, for example), the European Parliament (2021) stated that the data protection enforcement remains inconsistent or even absent across Member States, and concluded that enforcement had not significantly improved compared to the previous Data Protection Directive. Moreover, even though the GDPR as a regulation was adopted to harmonise the legal rules, legal scholars observe that data protection authorities (DPAs) apply varying levels and strategies of enforcement (Gentile & Lynskey, 2022; Sivan-Sevilla, 2022).

As was also discussed in Paper III, most DPAs have explicitly stated that they lack the human, technical, and financial resources, as well as adequate premises and infrastructure needed to effectively carry out their duties and exercise their powers (European Parliament, 2021). Again, this situation persists despite the Member States' obligation to sufficiently equip DPAs to enforce the GDPR. The issue of lack of genuine independence of DPAs in many Member States has been pointed out as well (Gentile & Lynskey, 2022; Veale & Borgesius, 2021).

A more recent assessment of the GDPR's enforcement paints a disappointing picture of enforcement level of the GDPR as well. An analysis conducted in January 2025 by an Austrian NGO NOYB concerning national DPAs enforcement activity — seven years after the GDPR came into effect — reveals that, on average, only 1.3% of cases handled by DPAs result in a fine (nyob, 2025a). This is particularly concerning given the strong enforcement tools available to DPAs, and in light of NOYB's survey which shows that it is the monetary fines that are the most significant factor motivating companies to comply with the rules[58]. Despite this, GDPR's enforcement remains weak, and as NYOB's founder Max Schrems put it,

[58] A survey that nyob has conducted among data protection professionals show that 67.4% of respondents view fines against their own company as a strong motivator for compliance, 61.5% also recognize the deterrent value of observing sanctions imposed on others, highlighting the broader signaling power of regulatory action (nyob, 2025a).

'[a]t the moment, DPAs often seem to be acting in the interests of companies rather than the people concerned' (nyob, 2025a).

Moreover, adding to the problem of effective enforcement of the GDPR is the emergence of the large language models (LLMs), which has triggered new challenges for the personal data protection in the EU. One of them concerns the problem of the so-called *hallucinations*[59] which sometimes occur during the operation of LLM models[60]. However, the way many companies develop these technologies often conflicts directly with other key provisions of the GDPR, as the datasets used for training, developing, and operating AI models frequently contain personal data. This has led several European DPAs to open investigations under the GDPR into how, for example, OpenAI handles personal data within the service. The European Data Protection Board (EDPB) has issued a Report of the work undertaken by the ChatGPT Taskforce (EDPB, 2024b, p. 5), in which it states:

> (…) controllers processing personal data in the context of LLMs shall take all necessary steps to ensure full compliance with the requirements of the GDPR. In particular, technical impossibility cannot be invoked to justify non-compliance with these requirements.

At the same time, the EDPB (2024) stated in another document[61] concerning the guidelines for the DPAs in interpreting some of the GDPR's rules in the context of AI models, that it 'generally recalls that [DPAs] enjoy discretionary powers to assess the possible infringement(s) and choose appropriate, necessary, and proportionate measures, taking into account the circumstances of each individual case'. In effect, the manner in which

[59] Hallucinations refer to instances in which the model generates false or fictional content, diverging from factual accuracy and potentially producing outputs not grounded in the data on which it was trained (Perković et al., 2024).

[60] For example, in August 2024, ChatGPT presented a Norwegian citizen as a convicted criminal who murdered two of his children and attempted to murder his third son. Although the story is fake in the part of the murder, it included real elements of his personal life, such as the number, gender, approximate age of his children, and the name of his home town (BBC, 2025).

[61] In *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models,* the EDPB was requested to issue an opinion on matters relating to the processing of personal data in the context of AI models. The request concerned the following issues: (1) when and how an AI model can be considered as 'anonymous'; (2) how controllers can demonstrate the appropriateness of legitimate interest as a legal basis in the development and (3) deployment phases; and (4) what are the consequences of the unlawful processing of personal data in the development phase of an AI model on the subsequent processing or operation of the AI model (EDPB, 2024).

DPAs interpret the GDPR in response to challenges posed by the use of LLMs is likely to differ among EU Member States.

Concerning the DSA's implementation by the oversight bodies – since the DSA was not yet in force during the writing of Paper II, its enforcement could not be studied. However, after the paper was published, the EU Commission started its first enforcement actions in October 2023 by requesting information from selected VLOPs[62]. For example, the Commission has sent formal requests for information to 17 of the VLOPs and VLOSEs seeking clarification on the measures they have implemented to comply with the obligation to grant eligible researchers timely access to publicly accessible data available through their online interfaces (European Commission, 2024e). Other requests concern information on the mitigation measures for risks linked to generative AI, such as the mentioned above 'hallucinations', the viral dissemination of deepfakes, as well as the automated manipulation of services that can mislead voters (European Commission, 2024d). Moreover, the ongoing Commission's investigations concern X, TikTok[63], AliExpress[64], Facebook and Instagram[65]. For instance, the proceedings initiated by the European Commission (2024a) address the issue of Meta's compliance with DSA obligations on the following:

> Assessment and mitigation of risks caused by the design of Facebook's and Instagram's online interfaces, which may exploit the weaknesses and inexperience of minors and cause addictive behaviour, and/or reinforce so-called 'rabbit hole' effect. Such an assessment is required to counter potential risks for the exercise of the fundamental right to the physical and mental well-being of children as well as to the respect of their rights.

[62] An overview of the main enforcement activities concerning the designated VLOPs and VLOSEs is provided on the Commission's webpage: https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses (accessed 10/03/2025).

[63] Regarding transparency and giving data access to researchers harmful content and addictive design but also for protection of minors (European Commission, 2024a).

[64] The investigations concern the lack of enforcement of AliExpress' terms of service prohibiting certain products posing risks for consumers' health (such as fake medicines and food as well as dietary supplements) and for minors specifically (access to pornographic material), which consumer can still find on the platform (European Commission, 2024f).

[65] Regarding deceptive advertisements and disinformation, visibility of political content, and 'the non-availability of an effective third-party real-time civic discourse and election-monitoring tool ahead of the upcoming elections to the European Parliament and other elections in various Member States', as well as the mechanism to flag illegal content.

With regard to the AI Act's enforcement framework, as discussed in Paper III, the similarities with GDPR's enforcement model suggest that similar concerns may emerge during AI Act's enforcement. As has been also pointed out in Paper IV, the potential risks may concern such issues as the uneven or varying implementation approaches, and insufficient resources of the oversight authorities. An additional complicating factor in terms of resourcing is that, depending on the national arrangements, the enforcement tasks may be divided between multiple MSAs. Member States must therefore ensure a clear allocation of tasks to maintain effective implementation of the AI Act. Given this challenge, engaging with experts from the established under the AIA *scientific panel* could potentially provide useful way for national authorities to address, for example, the technical challenges with scrutinising AI systems' compliance.

Moreover, it should be noted that the MSAs under the AI Act do not have the guaranteed status of independence as in the case of the DPAs. This concern may be even more amplified under the AI Act, as many high-risk AI systems will be used by the public sector. In particular, AI tools like remote biometric identification and AI used in the context of law enforcement, may increase the potential for abuse by governments. Although the AI Act contains safeguards to prevent such misuse, their effectiveness may be undermined in Member States lacking robust democratic checks and institutional accountability. [66]

5.2.3 AI transparency as a governance system

In the preceding analysis, AI transparency has been explored both as a governance ideal and as a governance tool within the framework of Trustworthy AI. The focus was on exploring how AI transparency has been conceptualised, designed, and implemented for individuals and oversight bodies. What becomes apparent is that AI transparency measures for these groups do not function in isolation. Rather, they operate in a broader,

[66] For example, Hungary's parliament has passed a law banning Pride events and permitting police use of facial recognition to identify participants, intensifying government measures against the LGBTQ+ community (Euronews, 2025c).

interconnected system of relations — in what may be called a governance system.

The theoretical perspectives presented in Chapter 3 and in Paper II offer helpful tools to understand this system, particularly through the work of Heald (2006) and Koivisto (2022). The transparency measures in the AI governance framework follow not only different transparency directions (as explored in Paper II and illustrated in Figure 3 below), but also are based on different transparency rationalities.

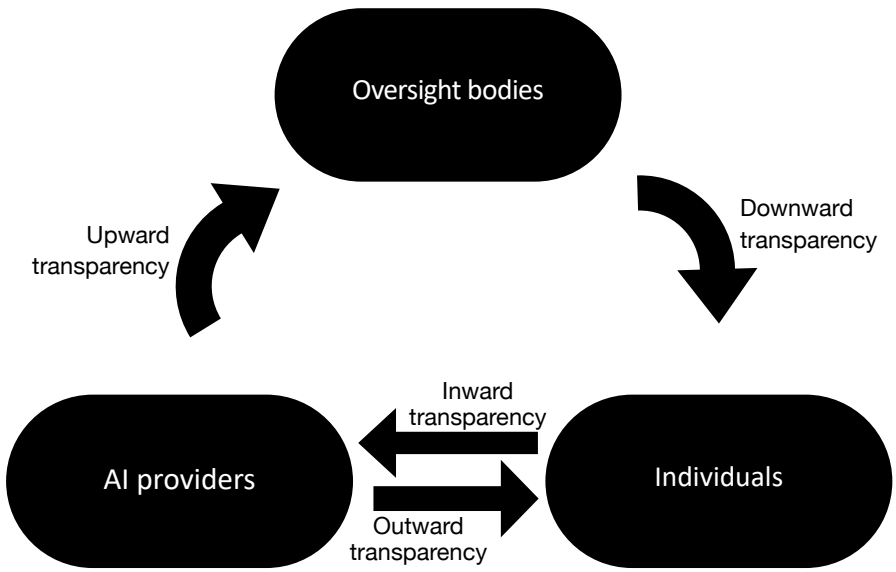

Fig. 3: Transparency directions in a governance system.

Most AI transparency measures within the GDPR, DSA, and AIA rely on the *public law rationality* of transparency (Koivisto, 2022). For example, individuals are entitled to receive certain information about AI systems (the inward transparency direction). However, as discussed in Paper I, this information is often incomplete and ambiguous. In any case, individuals do not have the legal ground to directly access the information held by AI providers to verify whether the information provided is complete or true. And arguably, they should not be expected to do so.

Instead, the legal mandates for conducting such checks are vested in the oversight bodies, what is represented as *upward transparency* in the model. Following this rationality of transparency, oversight bodies (as the governors) hold AI providers (the governed) accountable by ensuring system compliance and legal conformity. Conversely, following the *democratic rationality of transparency*, the downward transparency would allow individuals (as principals/governed) to hold the oversight bodies (as agents/governor) accountable for the oversight work in their stead.

In a well-functioning governance model of AI technologies, both modes of transparency (democratic and public law rationality) should be interwoven, as both are important to balance the power asymmetries between the state, powerful corporations, and citizens. Within such governance system, the various AI transparency directions and levels should interact, complement and reinforce each other.

The notion of a structured AI governance system — what Knowles and Richards (2021) call the 'systemness' of the framework— is central to their argument that public trust in AI depends on institutional authority. They stress the need to build a governance ecosystem with clearly defined rules for AI providers and mechanisms for enforcement to 'weed-out' AI systems that are harmful for societies. Grounding their argument in sociological theory, they point to the recursive relationship between institutional structures and the practices that shape them, arguing as follows:

> Because the effort required to forge and manage interpersonal trust relationships does not scale to the level of social complexity in these societies, a different basis for social order emerges from 'system trust' (i.e., trust in the functioning of bureaucratic sanctions and safeguards, especially the legal system).

That is to say, the objective of Trustworthy AI cannot be supported by transparency measures — such as requiring AI providers to supply information directly to users – in isolation. Rather, public trust in AI is forged through a well-functioning governance system as a whole, that is capable of ensuring that only AI systems that are lawful, ethical, and robust are deployed in the EU. The legal framework governing AI should therefore integrate mechanisms to create conditions for organisations to develop trustworthy AI and to employ them responsibly. When such conditions are absent, and untrustworthy AI systems are allowed to proliferate, there has

been an institutional failure to effectively govern AI (Knowles & Richards, 2021).

Thus, in light of the above analysis, it might be therefore asked: with all the legal concepts and requirements largely specified, could it be concluded that everything is in place to foster the objective of Trustworthy AI? The answer is not straightforward, as even the best laws on paper will not be realised unless they are implemented and enforced in alignment with the underlying governance ideals. Thus, what remains uncertain — and crucial — is how the above transparency governance tools will be operationalised.

5.3 AI transparency as a 'floating signifier'

In this section, I address the fourth and final meaning of AI transparency – as a 'floating signifier'. In Chapter 3, drawing on the work of Koivisto (2022), I have presented transparency as a concept that can be understood in many ways. The visual metaphor underlying transparency makes it a highly suggestive, versatile and malleable concept, allowing it to be framed in various ways depending on the context. As shown, the metaphor of transparency has over the past decades gained significant traction across economic, organisational, and political governance contexts, becoming closely associated with a governance ideal. Transparency has also emerged as a prominent socio-legal ideal in law, underpinning concepts such as rule-governed administration and public governance (Koivisto, 2022). Its appeal lies not only in its visual and cognitive metaphorical meaning – but also in the important values and objectives that it *enables*, such as meaningful accountability mechanisms. At the same time, the widespread popularity and positive associations with transparency may also lead to its semantic drift – when transparency is invoked in non-determinate ways, often as a rhetorical device, devoid of its inherent pro-ethical function (Turilli & Floridi, 2009).

Drawing parallels with the metaphorical meanings of transparency as a concept, it could also be argued that a similar phenomenon is observed with regard to the concept of *AI transparency* – with both 'AI' and 'transparency' being highly attractive and malleable concepts. It is therefore worth recognising that when we encounter claims asserting that a given AI system

is transparent, this does not necessarily mean that the system is transparent in the sense of complying with EU legal frameworks or with the Trustworthy AI guidelines. For example, information explaining transparency policy on companies' websites may often state as follows:

> The public and policy makers want to be better informed about our actions and we recognize these calls for greater transparency. That is why our original report has evolved into a more comprehensive X Transparency Center covering a broader array of our transparency efforts. We now include sections covering information requests, removal requests, copyright notices, trademark notices, email security, X Rules enforcement, platform manipulation, and state-backed information operations. (X, 2025)

Does it mean that X's use of recommender systems is aligned with the Trustworthy AI objective? In any case, it is not necessarily because the company claims so. As mentioned, in January 2025, the Commission has initiated investigative measures relating to the platform's recommender systems (European Commission, 2025b). Moreover, the new transparency path created by the DSA – the access to data by vetted researchers – is highly resisted by the platforms, which may be unsurprising. The Commission has initiated investigations concerning X for failing to provide access to its public data to researchers in line with the conditions set out in the DSA. As the Commission states, X prohibits vetted researchers from independently accessing its public data, such as by scraping. In addition, X's process to grant vetted researchers access to its application programming interface (API) appears to dissuade researchers from carrying out their research projects or leave them with no other choice than to pay disproportionally high fees (European Commission, 2024c).

Another example – Meta claims that it is transparent to the users (as in the previously mentioned privacy policies, and DSA reports). Below is an excerpt from META's newsroom post concerning 'New Features and Additional Transparency Measures as the Digital Services Act Comes Into Effect'[67]:

> We welcome the ambition for greater transparency, accountability and user empowerment that sits at the heart of regulations like the DSA, GDPR, and the ePrivacy Directive. […] We were the first platform to put in place ads transparency

[67] META. (2023). New Features and Additional Transparency Measures as the Digital Services Act Comes Into Effect | Meta. Https://About.Fb.Com/News/2023/08/New-Features-and-Additional-Transparency-Measures-as-the-Digital-Services-Act-Comes-into-Effect/. Accessed 10/03/2025

> tools and, for many years, we've provided industry-leading transparency for social issue, electoral and political ads (Meta, 2023).

At the same time, the European Commission (2024a) has started the investigations concerning Meta's deceptive advertisements and disinformation, visibility of political content, as well as the mechanism to flag illegal content. The European Commission (2024a) has noted, moreover, that Meta is in the process of deprecating 'a public insights tool that enables real-time election-monitoring by researchers, journalists and civil society', without an adequate replacement.

Although the investigations are ongoing at the time of writing, the key point to be made is that the technology companies clearly use the transparency narrative by publishing multiple transparency reports to create a public appeal of being genuinely transparent and trustworthy. As Koivisto (2022) notes, powerful technology actors have increasingly adopted the language of transparency to enhance their institutional legitimacy, although they largely operate beyond the reach of meaningful public oversight. Instead, information made available publicly reflects what companies *choose* to disclose. In such contexts, the so-called transparency may therefore be seen as a legitimising strategy, while not enabling the meaningful transparency tools to verify the provided information. Thus, the notion of AI transparency functions in the AI governance discourses as a 'floating signifier' as well, with the transparency narrative aligning more closely with corporate objectives and convenience than with legal requirements or informational needs of individuals.

6 Discussion: AI transparency serving the Trustworthy AI objective?

This chapter reflects on the findings of the thesis in light of its overarching aim: to contribute to a clarified understanding of AI transparency in the EU's governance framework, particularly in relation to the objective of Trustworthy AI. Drawing on the legal, conceptual and empirical analysis developed in Chapter 5 and in Papers I–IV, the chapter discusses how transparency has been conceptualised, designed, and implemented in the GDPR, the DSA, and the AIA for individuals and oversight bodies.

The chapter discusses the two guiding research questions and draws on the four conceptual dimensions of AI transparency developed in this thesis: transparency as a stand-alone objective, as a governance ideal, as a governance tool, and as a 'floating signifier'. Section 6.1 briefly returns to the above dimensions of AI transparency. Sections 6.2 and 6.3 discuss the analysis of how AI transparency has been envisioned, articulated and operationalised in EU's AI governance with regard to individuals and oversight bodies. Section 6.4 reflects how these findings relate to the evolving concept of Trustworthy AI. The chapter concludes in Section 6.5 by discussing the shifting political landscape of the EU, as it has unfolded during the final months of writing this thesis.

6.1 AI transparency as a multi-dimensional concept

As the analysis demonstrates, the four dimensions of AI transparency, outlined in Chapter 3 and expanded in Chapter 5, are all present in the EU's governance discourses concerning AI. Together, they can be seen as interrelated conceptual layers.

First, AI transparency has been presented as a stand-alone objective. It is grounded in the notion of an ideal — promising access to the 'truth' about AI systems, both in terms of their internal operations and their effects on

surrounding environments. However, as discussed in Section 5.1, the complexity of AI systems and the opacity inherent in many machine learning models constrain the degree to which such 'truth' can be accessible or meaningful. Still, transparency as a stand-alone objective may be seen as a heuristic approach to advance understanding of the workings of AI systems, their impacts, encouraging additional research, examinations, and testing.

Second, AI transparency has been presented as a governance ideal – an approach taking into account knowable aspects of AI systems, managing the 'unknowables', and legitimate interests of other stakeholders. Yet, due to the competing interests, limitations, and various needs of different AI stakeholder groups, the balance between transparency and secrecy is struck in different ways depending on the context. Still, it could represent the most optimal – perhaps unattainable – ideal of balance between all legitimate rights and interests, operationalized in a well-functioning, robust governance framework.

Third, AI transparency has been presented as a governance tool — that is, as legal translations intended to operationalise the AI transparency as a governance ideal outlined above. As shown, these technocratic translations should be designed in ways that enable the fulfilment of transparency objectives tailored to specific stakeholder groups and contexts. I have examined how such transparency tools have been designed and implemented across the GDPR, the DSA, and the AIA for two stakeholder groups: individuals and oversight bodies.

Fourth, I have discussed the concept of AI transparency operating as a 'floating signifier'. I have shown that the term draws on the metaphor of visibility and understanding, as well as on the normative appeal of transparency as a governance ideal, while remaining detached from clear legal obligations or accountability mechanisms. As such, AI transparency can be seen as used for rhetorical or legitimising purposes, and potentially lead to devaluation of AI transparency as a governance ideal.

In a way, the relations between the above dimensions of AI transparency could be compared to the broken telephone game, or illustrated by Plato's cave metaphor – yet modified by a few added layers of distortion. That is, if AI transparency as a *stand-alone objective* can be seen as an ideal, then a *governance ideal* can be seen as its imperfect representation in the real-

world contexts. Then, the governance ideal is further translated into a *governance tool* – such as legal rules – when more information may get lost or altered. Subsequently, such legal translations of AI transparency can again be implemented and enforced in different ways across jurisdictions, introducing additional possibilities for distortion. In the end, AI transparency that operates as a '*floating signifier'* may come to diverge so far from the original that it no longer bears a meaningful resemblance to the AI transparency ideal.

6.2 Limited AI transparency for individuals

The first research question addressed how AI transparency has been conceptualised, designed, and implemented for individuals. On the conceptual level — as a governance ideal — AI transparency has been framed as a way to help individuals broadly understand how AI systems function, including their capabilities and limitations. In the online environments, this governance ideal also includes providing individuals with tools to understand and 'fully control' how their personal data is processed, and to ensure their safety online.

However, the analysis in Chapter 5 and in Papers I-II shows that individual-oriented transparency remains limited in both scope and effect, often not providing the possibility of meaningful comprehension or control. The inherent limitations of individual-oriented AI transparency are present already at the conceptual level of the governance ideal. Such limitations include potential conflicts with other rights or interests, cognitive burden placed on individuals, limited possibility to create systemic change, and enforce accountability measures.

It has been shown that AI transparency which individuals may have access to in practice is the information about AI systems that is *communicated* to them by AI providers. Even if the right to explanation – as articulated in Art. 22 GDPR and Art. 86 AIA – is implemented in line with the governance ideal, the legal opacity barriers such as trade secrecy and business confidentiality would still apply. The empirical study presented in Paper I can be seen as an illustration of this. The Paper's findings show that data controllers provide vague and incomplete responses about the ADMs they use, often referring

to trade secrecy or business confidentiality limitations. However, the CJEU's rulings in *SCHUFA Holding* (C-634/21) and more recently in *CK v Dun & Bradstreet Austria* (C-203/22)(CJEU, 2025), have confirmed that the right to explanation in Article 22 applies in credit loan contexts, albeit still on case-by-case basis. While the right to explanation of automated decisions in scope of Art. 22 GDPR is still contested, and similar issues may be expected to arise with regard to the interpretation of Art. 86 AIA, such case law developments show promise that individuals may be provided with more meaningful AI transparency in certain contexts.

Still, the information made available to individuals is not necessarily truthful, fair, or non-discriminatory. Due to legal opacity barriers, individuals cannot directly hold AI providers accountable for such issues. Furthermore, as highlighted in the literature, it is unrealistic to expect individuals to fully understand or act on all the information they receive, given the human cognitive limitations. Understanding how AI systems were trained, how they are operating and what high-reach impact they may have on societies is beyond individual's reach and control as well.

Instead, it is the oversight bodies that have been entrusted with the responsibility to ensure that only *lawful*, *ethical,* and *robust* AI systems are the deployed within the Union. The ongoing operationalising of governance framework of Trustworthy AI will therefore be shaped primarily by the oversight bodies, national governments and EU institutions. Still, the individuals may – and should – hold these institutions accountable for the way their rights and interests are safeguarded, by utilising the democratic rationality of transparency.

6.3 In the authorities we trust?

While AI transparency tools that individuals have at their disposal are subject to inherent limitations – in contrast, oversight-oriented AI transparency represents the most comprehensive form of AI transparency available to third parties. Unlike transparency measures for other stakeholders, this level of transparency has been envisioned to allow the oversight authorities to effectively scrutinise the compliance of AI systems with relevant laws – including individuals' rights such as privacy, data

protection, non-discrimination, and safety. AI transparency for oversight purposes may thus be seen as playing a key role in maintaining the overall effectiveness of the EU legal frameworks.

The balance between transparency and secrecy interests is struck in a different way on this stakeholder level, as the trade secrecy claims are generally seen as not applicable in investigation contexts (e.g. de Laat, 2018). Transparency mandates for oversight bodies – what has been referred to as *qualified transparency* – are needed to verify whether the secrecy claims by AI providers are legitimate – or whether the secrecy laws are used to conceal illegitimate practices.

However, at the conceptual level of governance tool, the oversight-oriented AI transparency faces a number of structural and institutional challenges. While all three frameworks examined in this thesis grant oversight bodies broad mandates to access information about the design, functioning, and deployment of AI systems, legal mandates alone are not sufficient. Effective oversight depends on the way these institutions utilise their enforcement powers, resources, and on whether they exercise such tasks in an independent way from political and economic pressures.

Paper III illustrates that enforcement of the GDPR, which primarily relies on a decentralised enforcement pattern, has proven to be weak across many Member States, even years after the regulation became applicable. This weakness may reflect a broader trend in which Member States seek to attract AI businesses by limiting the regulatory burden.

This is concerning in view of the fact that the AIA is largely based on a similar enforcement pattern. As discussed in Paper IV, one of the potential enforcement challenges is the vast range of enforcement responsibilities of the MSAs at the national level. The oversight tasks are primarily designed to be exercised after the AI systems' deployment, with only limited scrutiny by notified bodies prior to the placement of high-risk AI systems on the EU market. Additional challenges stem from the technical complexity of many AI systems, which may exceed the capacity of some authorities. Although the AIA provides the possibility of access to the experts from the scientific panel, this mechanism is optional for the authorities. Enforcement of the AIA may be further complicated by the division of tasks among national authorities, which risks diluting responsibility and consistency in the AIA

interpretation and enforcement. Moreover, the issue of (non-) independence has been pointed out as potentially particularly worrying in Member States where democratic checks and balances are weakened.

By contrast, the relatively recent developments in the DSA's enforcement provides a more promising picture in this respect, as the Commission is proactively investigating a few major tech platforms with regard to their compliance with the new rules. This might suggest that the EU law enforcement is more effective under the centralised enforcement pattern. Yet, as will be elaborated below, the consistency and robustness of enforcement remain contingent on political and institutional alignment at the national and the EU level.

6.4 Trustworthy AI as a 'moving target'

Although the idea of Trustworthy AI has been outlined in the EU policy documents – primarily the Ethics Guidelines for Trustworthy AI – the interpretation of the concept in practice is not possible to determine. Moreover, as mentioned in Paper IV and above, the legal mandate of qualified transparency granted to oversight bodies does not automatically lead to investigation and enforcement measures. The way the Trustworthy AI principles will be interpreted are also likely to differ across jurisdictions and contexts. Thus, although the EU legal frameworks establish foundations for the Trustworthy AI, their adoption is merely the first step on the journey towards this objective.

In light of this, it could be argued that the concept of Trustworthy AI may also be shaped according to national and EU's political goals. On the one hand, the malleability of the concept can be seen as its advantage. It could be interpreted as 'a moving target' – a notion that evolves and progresses along with AI and societal developments, and their content can be updated whenever new risks, uses and challenges with AI arise (on the challenges and trade-offs between legal certainty and regulatory flexibility, see Larsson et al. 2024).

On the other hand, the way the concept of Trustworthy AI is interpreted by the policymakers, governments, and oversight bodies, may dictate *what*

Trustworthy AI is. In other words, the question which or whose purposes and interests the established governance framework will serve remains ultimately a political question. As seen in the analysis above on the oversight-oriented transparency tools and the transparency directions, much depends on the way oversight bodies will exercise their oversight tasks. Since AI transparency in the sense of full access to information for individuals is blocked *by design*, all the individuals are left with is *trusting* the oversight bodies. At the same time, the content of the Trustworthy AI may be filled with interests which are not fully aligned with those of the EU citizens. In light of such structural limitations of visibility and control over AI by the public, it could be argued that the concepts of AI transparency, as well as Trustworthy AI, may be used as a flexible tool to enable political or economic objectives. Will such objectives favour the welfare of EU citizens or other interests? Perhaps that will mean the conflation of trustworthiness with the acceptability of risks (Laux et al., 2024), which might invite further reflection.

Thus, just as transparency can function as a 'floating signifier' in AI governance discourses, the objective it serves — Trustworthy AI — may be subject to a similar process of semantic drift. Seen from such critical perspective, the concept of Trustworthy AI could be seen as joining terms like 'greenwashing' or 'transparency-washing' (Zalnieriute, 2021) as 'trustworthiness-washing'.

Further discussions should therefore be focused on exploring such questions as who will ultimately be the governor in the Trustworthy AI framework. Will it be EU citizens as the ultimate wielders of democratic power, and in whose interest the AI technology frameworks will be implemented and enforced?

6.5 The shifting political landscape

During the last months of writing this thesis, it appears that the political landscape surrounding AI governance in the EU has undergone a notable narrative shift. In late 2024, the so-called Mario Draghi (2024) report on European competitiveness advocated for 'simplifying rules' across the Union, warning that 'the stock of regulation remains large and new

regulation in the EU is growing faster than in other comparable economies'. This message was echoed by industry actors and political leaders who view complex regulation as a potential barrier to innovation, particularly in comparison to more permissive jurisdictions.

Across the Atlantic, in January 2025, the U.S. administration issued an executive order revoking President Biden's previous executive order from 2023 on the 'Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence', which established basic parameters for federal regulation and oversight of AI. The waves of AI deregulation have apparently reached the EU's shore as well, as in February 2025, on the AI Summit held in Paris, many of the talks by EU leaders seem to have adopted a similar tone. For example, French President stated, 'We are back in the race', adding that 'we will simplify' the rules, and that 'at the national and European scale, it is very clear that we have to resynchronize with the rest of the world' (NYT, 2025). In the same vein, Commissioner Henna Virkkunen, overseeing digital regulation, stressed that the AI Act must now be implemented in a 'very innovation-friendly manner', suggesting that reporting obligations under the Act might need to be revised or reduced. The forthcoming 'omnibus' legislative packages were described as tools for regulatory streamlining, aimed at recalibrating the EU's governance framework to reduce compliance burdens on industry (Euractiv, 2025).

One of the clearest signs of this changing political tone came with the Commission's withdrawal from its 2025 work agenda of the proposed AI Liability Directive. The Directive was intended to complement the AIA by providing harmonised rules for redress in cases of AI-related harm, yet according to the Commission, 'no foreseeable agreement' on the proposal was expected (Euronews, 2025a). The move sparked sharp criticism from key figures within the EU Parliament. German MEP Axel Voss called the decision a 'strategic mistake', while Brando Bonifei, co-rapporteur for the AI Act, described it as 'disappointing', stressing that harmonised liability rules would have provided much-needed clarity and fairness to both consumers and developers (Euronews, 2025d).

Although some have argued that AI liability issues could be addressed through the recently adopted Product Liability Directive (PLD), there is a significant difference between the two acts. The PLD applies to defective products and material damage, whereas the AI Liability Directive would

have addressed harms resulting from errors in algorithmic decision-making, including discriminatory outcomes. Crucially, it would have shifted the burden of proof onto developers in certain high-risk scenarios.

However, the Commission continues to express commitment to the original policy visions. In its 2025 *AI Continent Action Plan* (European Commission, 2025a), the Commission affirmed that AI developed and deployed in Europe must be 'safe, respect fundamental rights and is of the highest quality – a selling point for European providers – and drives the uptake of AI' — suggesting that regulatory objectives remain unchanged.

While supporting innovation and facilitating compliance are objectives that are commendable in themselves, these have been already designed within the Trustworthy AI approach. It is unclear if and how political sentiments may impact the implementation and enforcement of EU regulations such as the GDPR, DSA and AIA. While policy agility is important, it should not come at the expense of democratic accountability, legitimacy, and public trust. If AI is a race – not a cautious journey – the EU citizens are not likely to be the winners.

7 Conclusions

This thesis has examined the concept of AI transparency in the EU's evolving governance framework. By using a combination of legal-doctrinal and socio-legal approaches, I have explored the conceptual meanings and limitations of AI transparency across four levels of abstraction: as a stand-alone objective, as a governance ideal, as a governance tool, and as a 'floating signifier'. My particular focus has been on AI transparency understood as a governance ideal and as a governance tool in relation to the EU's policymaking objective of Trustworthy AI. These two meanings have guided my research questions, through which I have analysed how AI transparency has been conceptualised, designed, and implemented across the GDPR, the DSA, and the AIA, with regard to individuals and oversight bodies.

The thesis has shown that while individual-oriented AI transparency is framed as an important objective in the AI governance frameworks, its practical effect is often limited. As a governance ideal, AI transparency for individuals is meant to provide a general level of understanding of AI systems, control over personal data, and contribute to building public trust in the technologies. As a governance tool, it has been constructed as various rights – to notification, information or explanation. However, the findings of this thesis demonstrate that these rights are often vague in formulation, narrow in scope, and implemented in ways that limit their ability to support understanding and contestation.

By contrast, oversight-oriented AI transparency has been envisioned in the EU policy documents as a governance mechanism positioned to ensure that the important values, rights and interests of EU citizens are duly safeguarded, and that the stipulated laws are observed by AI providers. As a governance tool, AI transparency for oversight bodies emerges as comprehensive and far-reaching, at least on paper. The analysed legal frameworks – the GDPR, DSA and AIA – each introduce appropriate enforcement mechanisms, including investigation powers. Yet, the effectiveness of these tools remains dependent on the institutional capacity and independence from political or economic pressures.

At a conceptual level, the thesis has shown that AI transparency operates in the AI governance discourse in many dimensions. While the analysis has

primarily focused on exploring the concept as a governance ideal and a governance tool, the broader conceptual framing has opened up a discussion on the conceptual malleability of both AI transparency and Trustworthy AI — both of which risk becoming 'floating signifiers', drifting along with the changing political narratives.

In conclusion, the success of the EU's Trustworthy AI framework cannot be measured by its legal design alone. Much will depend on how transparency obligations are interpreted, implemented, and enforced in practice. Whether the governance ideal of AI transparency and the vision underpinning the Trustworthy AI will be realised — or lost in translation — remains a political and institutional question still in the making.

References

AI HLEG. (2019). Ethics guidelines for trustworthy AI. *Communication EC*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

algo:aware. (2018). *algo:aware Raising awareness on algorithms State-of-the-Art Report*. https://doi.org/https://actuary.eu/wp-content/uploads/2019/02/AlgoAware-State-of-the-Art-Report.pdf

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Ball, C. (2009). What Is Transparency? *Public Integrity*, *11*, 293–308. https://doi.org/10.2753/PIN1099-9922110400

Barocas, S., & Selbst, A. D. (2018). Big Data's Disparate Impact. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2477899

Baume, S., & Papadopoulos, Y. (2018). Transparency: from Bentham's inventory of virtuous effects to contemporary evidence-based scepticism. *Critical Review of International Social and Political Philosophy*, *21*(2). https://doi.org/10.1080/13698230.2015.1092656

BBC. (2025). *ChatGPT falsely told man he killed his children*. Https://Www.Bbc.Com/News/Articles/C0kgydkr516o.

Bird, C., Ungless, E., & Kasirzadeh, A. (2023). Typology of Risks of Generative Text-to-Image Models. *AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 396–410. https://doi.org/10.1145/3600211.3604722

Björverud, J. (2024). *Offentlighetsprincipen : Lagstiftarens avvägningar mellan handlingsoffentlighet och sekretess 1809-1980*. Lund University Publications.

Bradney, A. (2012). The Place of Empirical Legal Research in the Law School Curriculum. In *The Oxford Handbook of Empirical Legal Research*. https://doi.org/10.1093/oxfordhb/9780199542475.013.0043

Browne, R. (2025). *AI that can match humans at any task will be here in five to 10 years, Google DeepMind CEO says*.

https://www.cnbc.com/2025/03/17/human-level-ai-will-be-here-in-5-to-10-years-deepmind-ceo-says.html
Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*. https://doi.org/10.1177/2053951715622512
Burton, M. (2017). Doing empirical research exploring the decision-making of magistrates and juries. In *Research Methods in Law*. https://doi.org/10.4324/9781315386669
Busuioc, M. (2021). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review*, *81*(5). https://doi.org/10.1111/puar.13293
Busuioc, M., Curtin, D., & Almada, M. (2023). Reclaiming transparency: contesting the logics of secrecy within the AI Act. *European Law Open*, *2*(1), 79–105. https://doi.org/10.1017/ELO.2022.47
Calo, R. (2016). Robots as Legal Metaphors, 30 Harv. *Harvard Journal of Law & Technology*, *209*. https://digitalcommons.law.uw.edu/faculty-articles,https://digitalcommons.law.uw.edu/faculty-articles/18
Calo, R. (2017). *Artificial Intelligence Policy: A Primer and Roadmap*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3015350
Celliers, M., & Hattingh, M. (2020). A Systematic Review on Fake News Themes Reported in Literature. *Responsible Design, Implementation and Use of Information and Communication Technology*, *12067*, 223–234. https://doi.org/10.1007/978-3-030-45002-1_19/FIGURES/1
Chamberlain, J., & Kotsios, A. (2025). Defining Risk and Promoting Trust in AI Systems. In *EU Law in the Digital Age* (pp. 105–122). Hart Publishing. https://doi.org/10.5040/9781509981212.ch-007
CJEU. (2023). *SCHUFA Holding (Scoring) (C-634/21)*. https://doi.org/https://eur-lex.europa.eu/eli/C/2024/913/oj/eng
CJEU. (2025). *Dun & Bradstreet Austria (C-203/22)*. Https://Dpcuria.Eu/Case?Reference=C-203/22.
Cownie, F., & Bradney, A. (2017). Socio-legal studies a challenge to the doctrinal approach. In *Research Methods in Law*. https://doi.org/10.4324/9781315386669
Crawford, K. (2021). The Atlas of AI. In *The Atlas of AI*. https://doi.org/10.2307/j.ctv1ghv45t

Crawford, K., & Schultz, J. (2014). Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review*, *55*(93).

Cristianini, N. (2021). Shortcuts to Artificial Intelligence. In *Machines We Trust*. https://doi.org/10.7551/mitpress/12186.003.0006

de Laat, P. B. (2017). Big data and algorithmic decision-making. *ACM SIGCAS Computers and Society*, *47*(3), 39–53. https://doi.org/10.1145/3144592.3144597

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy and Technology*, *31*(4). https://doi.org/10.1007/s13347-017-0293-z

Dignum, V. (2019). Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. In *Arizona State Law Journal* (Vol. 51).

Directive 2024/2853. (n.d.). *Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC* (Directive (EU) 2024/2853). https://doi.org/http://data.europa.eu/eli/dir/2024/2853/oj

Draghi, M. (2024). *The Draghi report on EU competitiveness*. Https://Commission.Europa.Eu/Topics/Eu-Competitiveness/Draghi-Report_en.

Ebers, M. (2019). Chapter 2: Regulating AI and Robotics: Ethical and Legal Challenges. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3392379

EDPB. (2024a). *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models*. https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en

EDPB. (2024b). *Report of the work undertaken by the ChatGPT Taskforce*. https://doi.org/https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf

Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a Right to Explanation is Probably Not the Remedy You are Looking for. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2972855

Etzioni, A. (2010). Is Transparency the Best Disinfectant? *Journal of Political Philosophy*, *18*(4). https://doi.org/10.1111/j.1467-9760.2010.00366.x

Euractiv. (2025). *AI Summit: Regulation goes out of fashion in Paris*. Https://Www.Euractiv.Com/Section/Politics/News/Ai-Summit-Regulation-Goes-out-of-Fashion-in-Paris/.

Euronews. (2025a). *EU Commission to decide whether to scrap AI liability rules by August*. Https://Www.Euronews.Com/next/2025/03/18/Eu-Commission-to-Decide-Whether-to-Scrap-Ai-Liability-Rules-by-August.

Euronews. (2025b). *Hungary passes law banning Pride events in new blow to LGBTQ+ rights*. Https://Www.Euronews.Com/2025/03/18/Hungary-Passes-Law-Banning-Pride-Events-in-New-Blow-to-Lgbtq-Rights.

Euronews. (2025c). *Lawmakers reject Commission decision to scrap AI liability rules*. Https://Www.Euronews.Com/next/2025/02/18/Lawmakers-Reject-Commission-Decision-to-Scrap-Planned-Ai-Liability-Rules.

European Commission. (2010). A Digital Agenda for Europe. *Https://Eur-Lex.Europa.Eu/Legal-Content/EN/TXT/PDF/?Uri=CELEX:52010DC0245*.

European Commission. (2012a). Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). In *https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012PC0011*.

European Commission. (2012b). Safeguarding Privacy in a Connected World: A European Data Protection Framework for the 21st Century. In *https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012DC0009*.

European Commission. (2015). *A Digital Single Market Strategy for Europe*. Https://Eur-Lex.Europa.Eu/Legal-Content/EN/TXT/PDF/?Uri=CELEX:52015DC0192.

European Commission. (2017). *Communication from the Commission on the Mid-Term Review on the implementation of the Digital Single Market Strategy*. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52017DC0228

European Commission. (2018a). *Communication from the Commission on Artificial Intelligence for Europe.* https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237

European Commission. (2018b). *Communication from the Commission on Coordinated Plan on Artificial Intelligence*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0795

European Commission. (2020a). *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*. https://www.consilium.europa.eu/media/45910/021020-euco-final-conclusions.pdf

European Commission. (2020b). *Shaping Europe's digital future*. Https://Eur-Lex.Europa.Eu/Legal-Content/En/TXT/?Uri=CELEX%3A52020DC0067.

European Commission. (2020c). *White Paper on Artificial Intelligence: a European approach to excellence and trust*. https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Commission. (2021). *Impact assessment of the Digital Services Act | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/impact-assessment-digital-services-act

European Commission. (2023a). *Commission designates second set of VLOPs under the DSA*. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6763

European Commission. (2023b). *Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines [press release]*. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413

European Commission. (2024a). *Commission opens formal proceedings against Facebook and Instagram under the Digital Services Act*. https://doi.org/https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2373

European Commission. (2024b). *Commission requests information to Amazon under the Digital Services Act*. https://doi.org/https://digital-strategy.ec.europa.eu/en/news/commission-requests-information-amazon-under-digital-services-act

European Commission. (2024c). *Commission sends preliminary findings to X for breach of DSA*. https://doi.org/https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761
European Commission. (2024d). *Commission sends requests for information on generative AI risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines under the Digital Services Act*. https://doi.org/https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very
European Commission. (2024e). *Commission sends requests for information to 17 Very Large Online Platforms and Search Engines under the Digital Services Act*. https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-17-very-large-online-platforms-and-search-engines-under
European Commission. (2024f). *DSA: Commission opens formal proceedings against AliExpress*. https://doi.org/https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1485
European Commission. (2025a). *AI Continent Action Plan*. https://doi.org/https://commission.europa.eu/topics/eu-competitiveness/ai-continent_en
European Commission. (2025b). *Commission addresses additional investigatory measures to X in the ongoing proceedings under the Digital Services Act*.
European Council. (2017). *European Council meeting (19 October 2017) – Conclusions, EUCO 14/17*. https://data.consilium.europa.eu/doc/document/ST-14-2017-INIT/en/pdf
European Parliament. (2019a). *Digital Services Act and fundamental rights issues posed - Tuesday, 20 October 2020*. Https://Www.Europarl.Europa.Eu/Doceo/Document/TA-9-2020-0274_EN.Html.
European Parliament. (2019b). *Resolution on improving the functioning of the Single Market (2020/2018(INL))*.

https://doi.org/https://www.europarl.europa.eu/doceo/document/TA-9-2020-0272_EN.pdf
European Parliament. (2021). *Resolution on the Commission evaluation report on the implementation of the General Data Protection Regulation two years after its application*. https://www.europarl.europa.eu/doceo/document/TA-9-2021-0111_EN.html
Fenster, M. (2010). Seeing the state: Transparency as metaphor. *Administrative Law Review*, *62*(3).
Floridi, L. (2024). *Why the AI Hype is another Tech Bubble*. https://papers.ssrn.com/abstract=4960826
Forssbæck, J., & Oxelheim, L. (2014). *The Multifaceted Concept of Transparency*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199917693.013.0001
Fortes, P. R. B., Baquero, P. M., & Amariles, D. R. (2022). Artificial Intelligence Risks and Algorithmic Regulation. *European Journal of Risk Regulation*, *13*(3), 357–372. https://doi.org/10.1017/ERR.2022.14
Foss-Solbrekk, K. (2021). Three routes to protecting AI systems and their algorithms under IP law: The good, the bad and the ugly. *Journal of Intellectual Property Law & Practice*, *16*(3), 247–258. https://doi.org/10.1093/jiplp/jpab033
Foss-Solbrekk, K., & Glenster, A. K. (2022). The intersection of data protection rights and trade secret privileges in 'algorithmic transparency. In *Research Handbook on EU Data Protection Law*. https://doi.org/10.4337/9781800371682.00016
Fox, J. (2007). The uncertain relationship between transparency and accountability. *Development in Practice*, *17*(4–5), 663–671. https://doi.org/10.1080/09614520701469955
Freeman, K. (2016). Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis, 18 N.C. *NORTH CAROLINA JOURNAL OF LAW & TECHNOLOGY*, *18*, 12–13. https://scholarship.law.unc.edu/ncjolt
Galaz, V., Centeno, M. A., Callahan, P. W., Causevic, A., Patterson, T., Brass, I., Baum, S., Farber, D., Fischer, J., Garcia, D., McPhearson, T., Jimenez, D., King, B., Larcey, P., & Levy, K. (2021). Artificial

intelligence, systemic risks, and sustainability. *Technology in Society*, *67*, 101741. https://doi.org/10.1016/J.TECHSOC.2021.101741
Gentile, G., & Lynskey, O. (2022). DEFICIENT BY DESIGN? THE TRANSNATIONAL ENFORCEMENT OF THE GDPR. *International & Comparative Law Quarterly*, *71*(4), 799–830. https://doi.org/10.1017/S0020589322000355
Gibney, E. (2016). Google AI algorithm masters ancient game of Go. In *Nature* (Vol. 529, pp. 445–446). Nature Publishing Group. https://doi.org/10.1038/529445a
Hagendorff, T. (2024). Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. *Minds and Machines*, *34*, 39. https://doi.org/10.1007/s11023-024-09694-w
Halbertal, M. (2009). Concealment and revelation: Esotericism in Jewish thought and its philosophical implications. In *Concealment and Revelation: Esotericism in Jewish Thought and its Philosophical Implications*. https://doi.org/10.1163/187489109x12495426348843
Heald, D. A. (2006). *Varieties of transparency* (pp. 25–43). Oxford University Press. https://abdn.pure.elsevier.com/en/publications/varieties-of-transparency
Högberg, C. (2025). *"This ground truth is muddy anyway" : Ground truth data assemblages for medical AI development*. Sociologisk Forskning.
Hong, S., & Kim, S. H. (2016). Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*, *33*(4), 777–782. https://doi.org/10.1016/J.GIQ.2016.04.007
Hood, C. (2006). Transparency in historical perspective. In *Proceedings of the British Academy* (Vol. 135).
Hood, C. (2007). What happens when transparency meets blame-avoidance? *Public Management Review*, *9*(2), 191–210. https://doi.org/10.1080/14719030701340275
Hood, C. (2010). Accountability and transparency: Siamese twins, matching parts, awkward couple? *West European Politics*, *33*(5), 989–1009. https://doi.org/10.1080/01402382.2010.486122
Hutchinson, T. (2018). Doctrinal research, Researching the jury. In *Research Methods in Law*.

IEEE. (2022). IEEE Standard for Transparency of Autonomous Systems. *IEEE Std 7001-2021*, 1–54. https://doi.org/10.1109/IEEESTD.2022.9726144

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

Kaminski, M. E. (2020). Understanding Transparency in Algorithmic Accountability. In *The Cambridge Handbook of the Law of Algorithms* (pp. 121–138). Cambridge University Press. https://doi.org/10.1017/9781108680844.006

Kaminski, M. E. (2021). The right to explanation, explained. In *Research Handbook on Information Law and Governance*. https://doi.org/10.2139/ssrn.3196985

Kaminski, M. E. (2022). Regulating the Risks of AI. *Boston University Law Review*, *103*(5), 1347–1411. https://doi.org/10.2139/SSRN.4195066

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information Communication and Society*, *22*(14), 2081–2096. https://doi.org/10.1080/1369118X.2018.1477967

Knowles, B., & Richards, J. T. (2021). The sanction of authority: Promoting public trust in AI. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 262–271. https://doi.org/10.1145/3442188.3445890

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). A governance framework for algorithmic accountability and transparency. In *Eprs* (Issue April).

Koivisto, I. (2022). *The Transparency Paradox*. Oxford University Press. https://doi.org/10.1093/oso/9780192855466.001.0001

Koulu, R. (2021). Crafting Digital Transparency: Implementing Legal Values into Algorithmic Design. *Critical Analysis of Law*, *8*(1).

Koulu, R., Peters, A., & Pohle, J. (2021). Finding Design Patterns in Law: An Exploratory Approach. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3814234

Kronblad, C., Essén, A., & Mähring, M. (2024). When Justice is Blind to Algorithms: Multilayered Blackboxing of Algorithmic Decision Making in the Public Sector. *MIS Quarterly*. https://doi.org/10.25300/misq/2024/18251

Larsson, S. (2021). AI in the EU: Ethical Guidelines as a Governance Tool. In *The European Union and the Technology Shift*. https://doi.org/10.1007/978-3-030-63672-2_4

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2), 1–16. https://doi.org/10.14763/2020.2.1469

Larsson, S., Hildén, J., & Söderlund, K. (2024). *Between Regulatory Fixity and Flexibility in the EU AI Act*. Https://Portal.Research.Lu.Se/En/Publications/between-Regulatory-Fixity-and-Flexibility-in-the-Eu-Ai-Act.

Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, *18*(1), 3–32. https://doi.org/10.1111/REGO.12512

Leerssen, P. (2021). Platform research access in Article 31 of the Digital Services Act. *Verfassungsblog*. http://dx.doi.org/10.17176/20210907-214355-0

Leino-Sandberg, P. (2025). Who can guard the guardian? *European Law Blog*. https://doi.org/10.21428/9885764c.04117ca4

Lemley, M. A., & Casey, B. (2019). You Might Be a Robot. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3327602

Lessig, L. (1999). *Code and Other Laws of Cyberspace*.

Littman, M., Ajunwa, I., & Berger, G. (2021). *One Hundred Year Study on Artificial Intelligence (AI100)*. Stanford University, Stanford, CA. http://ai100.stanford.edu/2021-report.

Lomborg, S., Kaun, A., & Scott Hansen, S. (2023). Automated decision-making: Toward a people-centred approach. In *Sociology Compass* (Vol. 17). John Wiley and Sons Inc. https://doi.org/10.1111/soc4.13097

Lu, S. (2024). Regulating Algorithmic Harms. *Florida Law Review*. https://doi.org/10.2139/SSRN.4949052

Malabou, C. (2022). *Floating Signifiers Revisited: Poststructuralism Meets Neurolinguistics*. Https://Www.Cambridge.Org/Core/Books/Abs/Plasticity/Floating-Signifiers-Revisited-Poststructuralism-Meets-Neurolinguistics/699620C2942B4F56C7905EC47FE2C9B9.

Meijer, A. (2014). Transparency. In *The Oxford Handbook of Public Accountability:* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641253.013.0043

Meta. (n.d.). *Transparency reports | Transparency Center*. Https://Transparency.Meta.Com/Reports/.

Meta. (2023). *New Features and Additional Transparency Measures as the Digital Services Act Comes Into Effect | Meta*. Https://About.Fb.Com/News/2023/08/New-Features-and-Additional-Transparency-Measures-as-the-Digital-Services-Act-Comes-into-Effect/.

Mitchell, M. (2019). Artificial Intelligence: A Guide for Thinking Humans. Pinguin Books.

Mulder, W. De, Valcke, P., Vanderstichele, G., & Baeck, J. (2021). Are Judges More Transparent Than Black Boxes? A Scheme To Improve Judicial Decision-Making By Establishing A Relationship With Mathematical Function Maximization. *Law and Contemporary Problems*, *84*(3).

nyob. (2025a). *Data Protection Day: Only 1.3% of cases before EU DPAs result in a fine*. Https://Noyb.Eu/En/Data-Protection-Day-Only-13-Cases-Eu-Dpas-Result-Fine.

nyob. (2025b). *Swedbank refuses transparency in automatic interest calculation*. Https://Noyb.Eu/En/Swedbank-Refuses-Transparency-Automatic-Interest-Calculation.

NYT. (2025). *Macron Pitches Lighter Regulation to Fuel A.I. Boom in Europe*. Https://Www.Nytimes.Com/2025/02/10/Business/Ai-Summit-Paris.Html.

OECD. (2024). *Artificial Intelligence in Society*. OECD. https://doi.org/10.1787/EEDFEE77-EN

OECD. (2025). *Google DeepMind CEO Predicts Human-Level AI within 5-10 Years*. https://doi.org/https://oecd.ai/en/incidents/2025-03-17-6d72

Olsen, H. P., Hildebrandt, T. T., Wiesener, C., Larsen, M. S., & Ammitzbøll Flügge, A. W. (2024). The Right to Transparency in Public Governance: Freedom of Information and the Use of Artificial Intelligence by Public Agencies. *Digital Government: Research and Practice*, *5*. https://doi.org/10.1145/3632753

O'Neil, C. (2016). *Weapons of math destruction : how big data increases inequality and threatens democracy*. Pinguin Books.

O'Neil, O. (2002). A Question of Trust. Cambridge University Press.
Pasquale, F. (2010). Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries. *Northwestern University Law Review*, *104*(1), 105–174.
Pasquale, F. (2015). *The black box society : the secret algorithms that control money and information*. Harvard University Press.
Perković, G., Drobnjak, A., & Botički, I. (2024). Hallucinations in LLMs: Understanding and Addressing Challenges. *2024 47th ICT and Electronics Convention, MIPRO 2024 - Proceedings*, 2084–2088. https://doi.org/10.1109/MIPRO60963.2024.10569238
Poell, T., Nieborg, D., & van Dijck, J. (2019). Platformisation. *Internet Policy Review*, *8*. https://doi.org/10.14763/2019.4.1425
Regulation 2016/679. (n.d.). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. http://data.europa.eu/eli/reg/2016/679/oj
Regulation 2017/745. (n.d.). *Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices*. http://data.europa.eu/eli/reg/2017/745/oj
Regulation 2019/1020. (n.d.). *Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011*. http://data.europa.eu/eli/reg/2019/1020/oj
Regulation 2022/2065. (n.d.). *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. http://data.europa.eu/eli/reg/2022/2065/oj
Regulation 2024/1689. (n.d.). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. http://data.europa.eu/eli/reg/2024/1689/oj
Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Wagner Meira, W. M. (2020). Auditing radicalization pathways on YouTube. *FAT* 2020 -*

Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3351095.3372879

Richardson, R. (2022). *Defining and Demystifying Automated Decision Systems* . Https://Digitalcommons.Law.Umaryland.Edu/Cgi/Viewcontent.Cgi?Article=3930&context=mlr&utm_source=chatgpt.Com.

Ruschemeier, H. (2023). AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal. *ERA Forum*, *23*(3), 361–376. https://doi.org/10.1007/S12027-022-00725-6/METRICS

Russell, S., & Norvig, P. (2021). Artificial Intelligence - A Modern Approach. In *Pearson Series* ((4th Edition)).

Scherer, M. U. (2015). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2609777

Schudson, M. (2015). The Rise of the Right to Know. In *The Rise of the Right to Know*. https://doi.org/10.4159/9780674915787

Sebastian, G. (2023). Privacy and Data Protection in ChatGPT and Other AI Chatbots. *International Journal of Security and Privacy in Pervasive Computing*, *15*, 1–14. https://doi.org/10.4018/ijsppc.325475

Sivan-Sevilla, I. (2022). Varieties of enforcement strategies post-GDPR: a fuzzy-set qualitative comparative analysis (fsQCA) across data protection authorities. *Journal of European Public Policy*. https://doi.org/10.1080/13501763.2022.2147578

Söderlund, K. (n.d.). *High-risk AI transparency? On qualified transparency mandates for oversight bodies under the EU AI Act*.

Söderlund, K., Engström, E., Haresamudram, K., Larsson, S., & Strimling, P. (2024). Regulating high-reach AI: On transparency directions in the Digital Services Act. *Internet Policy Review*, *13*(1). https://doi.org/10.14763/2024.1.1746

Solove, D. J. (2024). Artificial Intelligence and Privacy. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4713111

Steimers, A., & Schneider, M. (2022). Sources of Risk of AI Systems. *International Journal of Environmental Research and Public Health*, *19*(6). https://doi.org/10.3390/IJERPH19063641

Taekema, S., & van der Burg, W. (2024). Introduction: Methodology of legal research. In *Contextualising Legal Research* (pp. 2–21). Edward Elgar Publishing. https://doi.org/10.4337/9781035307395.00008

Taylor, I. (2023). Justice by Algorithm: The Limits of AI in Criminal Sentencing. *Criminal Justice Ethics*, *42*(3), 193–213. https://doi.org/10.1080/0731129X.2023.2275967

Tryckfrihetsförordning (1949:105). (n.d.). Https://Www.Riksdagen.Se/Sv/Dokument-Och-Lagar/Dokument/Svensk-Forfattningssamling/Tryckfrihetsforordning-1949105_sfs-1949-105/#K2.

Tschider, C. (2021). Legal Opacity: Artificial Intelligence's Sticky Wicket. *Iowa Law Review*.

Turilli, M., & Floridi, L. (2009). The Ethics of Information Transparency. *Ethics and Information Technology*, *11*(2), 105–112. https://link.springer.com/article/10.1007/s10676-009-9187-9

van Bekkum, M., & Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security*, *23*(4), 323–340. https://doi.org/10.1177/13882627211031257/ASSET/IMAGES/LARGE/10.1177_13882627211031257-FIG2.JPEG

Veale, M., & Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112. https://doi.org/10.9785/cri-2021-220402

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2). https://doi.org/10.1093/idpl/ipx005

Wagner, M., Gupta, R., Borg, M., Engström, E., & Lysek, M. (2025). AI Act High-Risk Requirements Readiness: Industrial Perspectives and Case Company Insights. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* , *15453 LNCS*, 67–83. https://doi.org/10.1007/978-3-031-78392-0_5

Watkins, D., & Burton, M. (2018). *Research methods in law* (Second edition.). https://doi.org/10.4324/9781315386669

Westerman, P. C. (2011). Open or Autonomous? The Debate on Legal Methodology as a Reflection of the Debate on Law. In *Methodologies of Legal Research: Which Kind of Method for What Kind of Discipline?* https://doi.org/10.2139/ssrn.1609575

Wieringa, M. (2023). "Hey SyRI, tell me about algorithmic accountability": Lessons from a landmark case. *Data & Policy*, *5*, e2. https://doi.org/10.1017/DAP.2022.39

X. (2025). *X Transparency Center*. Https://Transparency.x.Com/En.

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information Communication and Society*, *20*(1). https://doi.org/10.1080/1369118X.2016.1186713

Yeung, K., & Ranchordás, S. (2024). *An Introduction to Law and Regulation*. Cambridge University Press. https://doi.org/10.1017/9781009379007

Zalnieriute, M. (2021). "Transparency Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism. *Critical Analysis of Law*, *8*, 139–153. https://doi.org/10.33137/cal.v8i1.36284

Zech, H. (2023). How Should We Regulate AI? *Weizenbaum Journal of the Digital Society*, *3*(3). https://doi.org/10.34669/WI.WJDS/3.3.7

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, *15*(11), e1002683. https://doi.org/10.1371/JOURNAL.PMED.1002683

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books.