



LUND UNIVERSITY

Transparency in artificial intelligence

Larsson, Stefan; Heintz, Fredrik

Published in:
Internet Policy Review

DOI:
[10.14763/2020.2.1469](https://doi.org/10.14763/2020.2.1469)

2020

Document Version:
Publisher's PDF, also known as Version of record

[Link to publication](#)

Citation for published version (APA):
Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, 9(2), 1-16.
<https://doi.org/10.14763/2020.2.1469>

Total number of authors:
2

Creative Commons License:
CC BY

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



Transparency in artificial intelligence

Stefan Larsson

Department of Technology and Society, Lund University, Sweden, stefan.larsson@lanth.lth.se

Fredrik Heintz

Department of Computer Science (IDA), Linköping University, Sweden

Published on 05 May 2020 | DOI: 10.14763/2020.2.1469

Abstract: This conceptual paper addresses the issues of transparency as linked to artificial intelligence (AI) from socio-legal and computer scientific perspectives. Firstly, we discuss the conceptual distinction between transparency in AI and algorithmic transparency, and argue for the wider concept ‘in AI’, as a partly contested albeit useful notion in relation to transparency. Secondly, we show that transparency as a general concept is multifaceted, and of widespread theoretical use in multiple disciplines over time, particularly since the 1990s. Still, it has had a resurgence in contemporary notions of AI governance, such as in the multitude of recently published ethics guidelines on AI. Thirdly, we discuss and show the relevance of the fact that transparency expresses a conceptual metaphor of more general significance, linked to knowing, bringing positive connotations that may have normative effects to regulatory debates. Finally, we draw a possible categorisation of aspects related to transparency in AI, or what we interchangeably call AI transparency, and argue for the need of developing a multidisciplinary understanding, in order to contribute to the governance of AI as applied on markets and in society.

Keywords: Transparency in AI, Algorithmic transparency, Explainable AI, AI governance

Article information

Received: 30 Apr 2019 **Reviewed:** 27 Sep 2019 **Published:** 05 May 2020

Licence: Creative Commons Attribution 3.0 Germany

Funding: The research for this paper has in part been funded by The Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS), within the AI Transparency and Consumer Trust project, as well as The Swedish Research Council (VR; grant no. 2019-00198) in the AIR Lund (Artificially Intelligent use of Registers at Lund University) research environment, and The Swedish Retail and Wholesale Council (DATA/TRUST 2018:787).

Competing interests: The author has declared that no competing interests exist that have influenced the text.

URL: <http://policyreview.info/concepts/transparency-artificial-intelligence>

Citation: Larsson, S. & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, 9(2). DOI: 10.14763/2020.2.1469

INTRODUCTION: TRANSPARENCY IN AI

Transparency is indeed a multifaceted concept used by various disciplines (Margetts, 2011; Hood, 2006). Recently, it has gone through a resurgence with regards to contemporary discourses around artificial intelligence (AI). For example, the ethical guidelines published by the EU Commission’s High-Level Expert Group on AI (AI HLEG) in April 2019 states transparency as one of seven key requirements for the realisation of ‘trustworthy AI’, which also has made its clear mark in the Commission’s white paper on AI, published in February 2020. In fact, “transparency” is the single most common, and one of the key five principles emphasised in the vast number – a recent study counted 84 – of ethical guidelines addressing AI on a global level (Jobin et al., 2019). Furthermore, there is a critical discourse on AI and machine learning about fairness, accountability and *transparency*.¹ The number of publications in the fields related to AI and machine learning combined with ethics, governance and norms have grown remarkably over the last 2-5 years (Larsson et al., 2019).

While our conceptual focus here is on transparency, an important qualifier is AI, which in combination is highly interrelated to algorithmic transparency. While algorithmic transparency and algorithmic decision-making have become accepted terminology in contemporary critical research, we see a need for a more nuanced and elaborated terminology in its relationship to AI - to be able to clarify the conceptual framing of transparency.

On the one hand, AI indeed is a contested concept that lacks clear consensus, both in computer science (Monett, Lewis & Thórisson, 2020), law (Martinez, 2019) and the public perception (Fast & Horvitz, 2017). This is linked to the fact that intelligence alone has been defined in at least 70 different ways (Legg & Hutter, 2007). Furthermore, the definition has changed as the possibilities within the field has developed since its inception in the 1950s, posing what sometimes is called the “AI effect” or an “odd paradox” (discussed by Stone et al., 2016; see also McCorduck & Cfe, 2004) in the sense that once a problem seen as needing AI has been solved, the application ceases to be perceived as intelligent. This corresponds to the view that AI is about solving problems that computers currently cannot do, and as soon as it is possible for a computer to solve it, it no longer counts as an AI-problem. So, the hard-to-define field of AI has fittingly been addressed as not a single technology, but rather “a set of techniques and sub-disciplines ranging from areas such as speech recognition and computer vision to attention and memory, to name just a few” (Gasser & Almeida, 2017, p. 59).

On the other hand, there is ambiguity also in the ‘algorithmic’ concept, although it seems far less problematised in critical research. Firstly, the notion of algorithms in computer science as a finite step-by-step description on how to solve a particular class of problems – and hence what ‘algorithmic’ transparency would denote – is arguably narrower than how the concept is used in literature on governance issues, often relating to issues of accountability. For example, a recent report on “algorithmic transparency” lists seven points of what needs to be addressed. Only one of these are aimed specifically at algorithms, while the other six deals with issues of data, goals, outcomes, compliance, influence, and usage (Koene et al., 2019). While all of these aspects are highly relevant from a governance perspective addressing issues of accountability in relation to transparency, this also speaks for the ambiguity of the use of “algorithmic” in relation to transparency. Is it addressing a specific technological aspect or a systemic quality?

In line with this, and in relation to issues of unfair outcomes of algorithmic systems, it is often concluded that the specific algorithms and the code are very unlikely intended to discriminate in

a harmful way (Bodo et al., 2017). The challenge is one of relations between data and algorithms, emergent properties of the machine learning process, very likely to be unidentifiable from a review of the code. This also means that it is important to consider the context of the combination of machine learning algorithms, the underlying training data and the decisions they inform (Kemper & Kolkman, 2019). So, a key question is for whom the AI-systems or algorithmic decision-making should be more transparent. This is highlighted in relation to digital platforms on a global scale (Larsson, 2019), and Kemper and Kolkman (2019) argue for the need of a “critical audience”. Pasquale (2015, pp. 160-165) has called for “qualitative transparency”, which Larsson (2018) has interpreted as a need for supervisory authorities to develop methods for algorithmic governance.

The multitude of aspects combined with the complexity of context leads us to argue for a more systemic approach, here signified by the AI concept, as a wider notion than ‘algorithmic’ (see Doshi-Velez et al., 2017; Floridi et al., 2018). Furthermore, a reason is to strengthen a conceptual bridge between the fields of research dealing with ‘algorithmic transparency’ and accountability, on the one hand, and the fields researching AI and challenges of transparency, albeit in terms of making models more explainable and interpretable, on the other (see Lepri et al., 2018). Of particular interest, here, is the relationship between transparency and trustworthy AI, which is a key objective for the European AI strategy from April 2018, which not the least is emphasised by the subsequent AI HLEG’s ethics guidelines on *trustworthy AI* (2019) and a clear part of the “ecosystem of trust” sought for in the Commission’s white paper on AI (2020, p. 9).

Even if research related to transparency in AI, or what we interchangeably call *AI transparency*, recently has been claimed to be “in its infancy” (Theodorou, Wortham, & Bryson, 2017) the theoretical backdrop of transparency is, as mentioned, in itself vast and rather complex. Therefore, some of that backdrop, with its multidisciplinary and historical connotations of the concept, will be further addressed in the following section (1). Transparency, we show, comes with a metaphorical framing we analyse in order to show normative connotations attached to it. Neighbouring concepts like openness and explainability lead us to propose a categorisation of aspects of relevance to transparency in AI in Section 2. These are of relevance for the ethical and legal challenges outlined in Section 3.

1. HISTORICAL AND CONCEPTUAL DEVELOPMENT OF TRANSPARENCY

In an essay from 2009, Carolyn Ball analyses the metaphorical content of transparency as it had developed in the anti-corruption work by NGOs and supranational institutions in the 1990s. She describes the academic interest seen in publications around transparency terminology. Consistent with this, a search on ‘transparency’ in the Web of Science – which mainly indexes articles published in international scientific journals – reveals that it is a relatively newfound concept in the sense that there is a steady increase in the use of the concept from the 1990s onwards, see [Figure 1](#) below.²

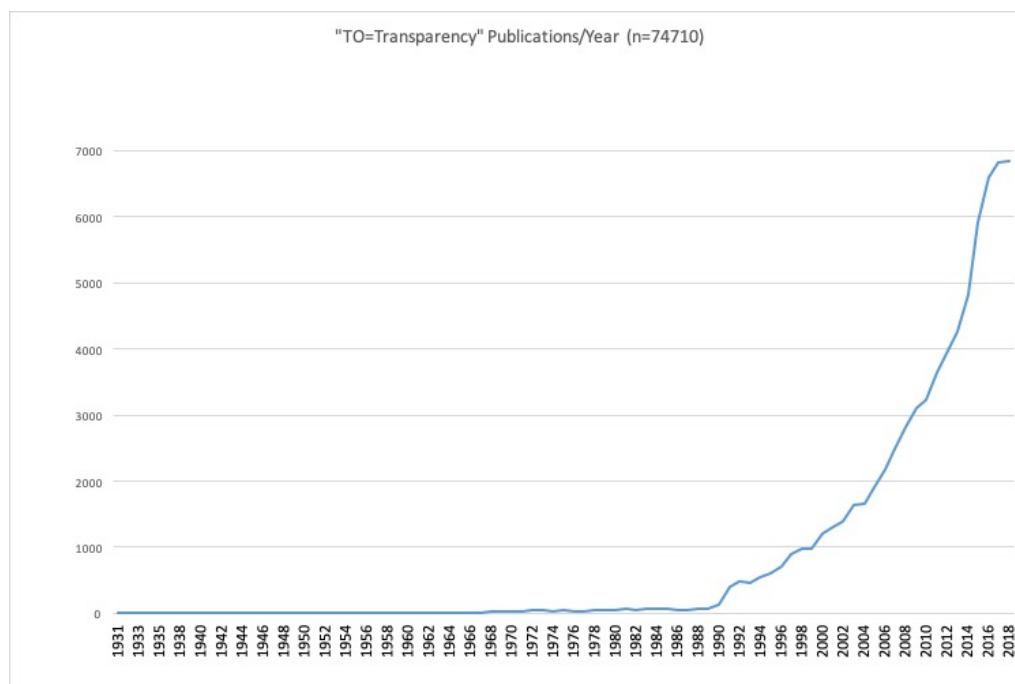


Figure 1: 'Transparency' as a concept in the Web of Science, 1931-2018.

Transparency has for example, according to Forssbæck and Oxelheim (2015), become a catchword in the economic-political debate, perhaps particularly in relation to a series of financial crises in the mid-1990s but also a series of scandals in the early 2000s leading to heightened interest in corporate governance. The EU's Transparency Directive from 2004 can be mentioned here. Linking transparency to economic theory, Forssbæck and Oxelheim tie the concept to the notion of information asymmetries, that is, where one party has more or better information than the other. This notion is also found in literature on fairness in algorithmic decision-making (Lepri et al., 2018).

One notable difficulty for theorising transparency, as pointed out by Hansen, Christensen and Flyverbom (2015, p. 118) has to do with the concept itself, and that it refers to such a wide array of objects, uses, technologies and practices. This is evident in a bibliometric overview of how the concept of 'transparency' is used in different research areas, see [Figure 2](#) below.

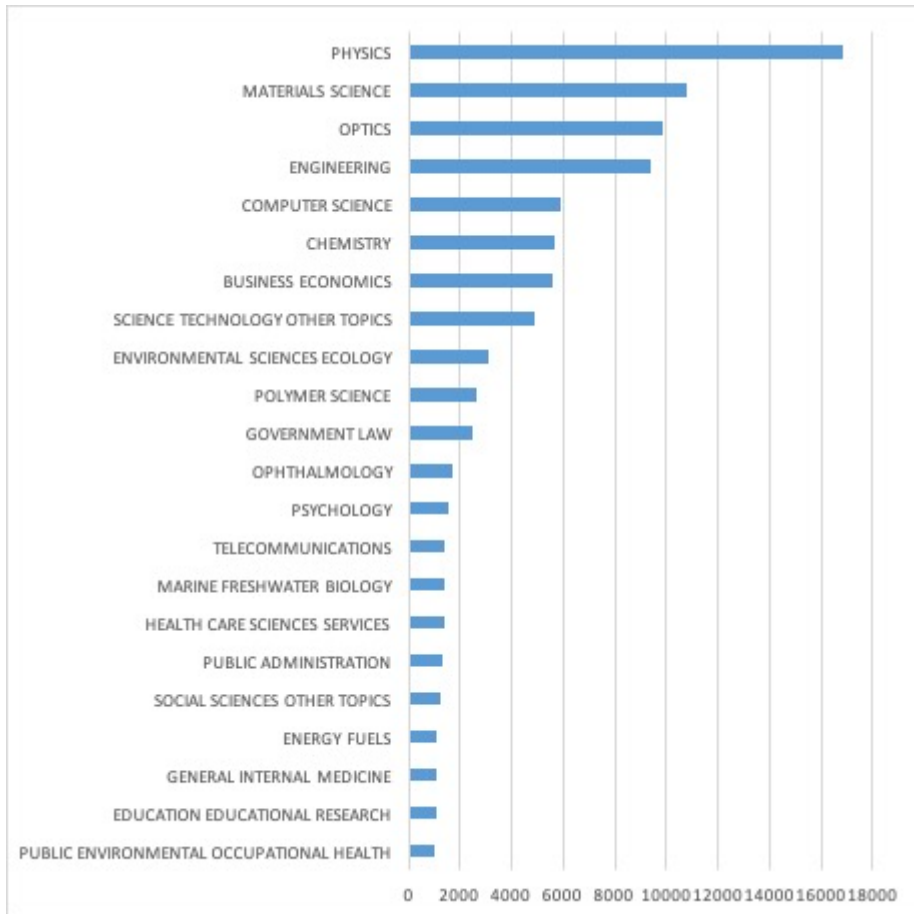


Figure 2: ‘Transparency’ use in different research areas, >1,000 publications, based on Web of Science journal classification categories.

The richness in the use of ‘transparency’ as a concept, as well as part of the difficulty to define it, relates to the fact that for some fields transparency denotes the physical property of a material and its capacity to allow light to pass through it, while in others it is thought of as a “powerful means towards some desirable social end, for example, holding public officials accountable, reducing fraud and fighting corruption” (Hansen, Christensen, & Flyverbom, 2015, p. 118). These different understandings should also be noted in relation to the uses of the concept in different disciplinary publications, see Figure 2.

THE METAPHORICAL FRAMING OF TRANSPARENCY

The conceptual metaphor knowing is seeing is depicted by Michael Reddy in 1979 (Reddy, 1979; see Larsson, 2017, p. 32). Reddy described the conduit metaphor system, a systemic set of mappings from the source domain of physical objects to the target domain of mental operations. This means that there are common metaphorical mappings for human understanding that structure aspects of knowledge to the metaphorical use of “seeing”. For example, when we conceptualise understanding in terms of seeing, which is commonplace, this also follows from a series of other closely linked expressions or associations that have to do with the condition *to see* (Lakoff & Johnson, 1980, 1999; Larsson, 2014). This includes light, brightness, clarity – and transparency. It is therefore likely hard to avoid this particular metaphorical mapping, thereby leading to a labelling of transparency linked to the positive labelling of knowledge as something good to have. That is, the benign conception of transparency relates to a deeper cognitive frame linked to knowing and understanding. And, conversely, the countering metaphors with negative connotations relates to being in the dark, perhaps most clearly displayed by the very much

present ‘black box’ terminology. The link between knowledge and transparency may however be illusive for particular contexts, as it also can be used for more rhetorical reasons, for example to deflect regulation (see Crain, 2018, below), or have unintended consequences.

There are a number of neighbouring as well as antonymic concepts of particular relevance for transparency as it relates to AI, such as ‘explainability’ (as in the research strand xAI), ‘black box’ (particularly popularised by legal scholar Frank Pasquale in *Black Box Society*, 2015) and ‘openness’. First of all, the clear metaphoricity of these concepts is relevant in itself for understanding the role and meaning of the terminology. The conceptual and metaphorical essence to the concept of transparency, and its theoretical implications is witnessed by Hansen, Christensen and Flyverbom (2015) as well as Christensen and Cornelissen (2015). Hansen, Christensen and Flyverbom (2015) address the normative challenge that many contemporary societal projects generally assume that transparency can effectively steer individual and collective behaviour towards desirable objectives.

The metaphorical analogy of a physical feature has, as argued by Koivisto (2016), led to that “transparency has come to denote a modern, surprisingly complex and expanding socio-legal ideal” – and therefore also has become a normative concept bearing premises that needs to be highlighted and discussed too. As a result, transparency’s negative connotations are, according to Koivisto, undertheorised. Ananny and Crawford (2018) revisits these general but metaphorically based notions of transparency in the context of algorithmic accountability. Their argument supports the notion of a wider transparency concept than what the narrower explainability domain focuses. It does so by rather than privileging a type of accountability that needs to look *inside* systems, instead hold systems accountable by looking *across* them: “seeing them as sociotechnical systems that do not *contain* complexity but *enact* complexity by connecting to and intertwining with assemblages of humans and non-humans” (Ananny & Crawford, 2018, p. 974). The embodiment of transparency is evident, in the sense that it structures our thinking. How AI and algorithmic systems are understood has normative effects on regulatory debates around how to govern AI.

NEIGHBOURING CONCEPTS – OPENNESS AND EXPLAINABILITY

As mentioned, ‘openness’ is tightly linked to transparency. It is a concept often framed with positive values, portrayed by ‘open data’, ‘open source’, ‘open code’ and ‘open access’ (Larsson, 2017, p. 215-220), as well as ‘open science’ (see Fecher & Friesike, 2014). Transparency in the sense of ‘open government’ comes with a political framing of empowerment under the notion of fostering democratic processes (see Ruijter, Grimmelikhuijsen, & Meijer, 2017). ‘Openness’ can also, for example, take place in a still ongoing battle between content-producing industries – traditionally relying on intellectual property regulation – and other industries relying on a freer flow of content, the so-called “openness industries” (Jakobsson, 2012; see Larsson, 2017). A challenge, addressed by Hansen, Christensen, and Flyverbom (2015) in terms of “transparency as paradox”, is that also a genuinely well-intended discourse of openness may lead to unintended consequences. For example, the transparency of social media platforms – mentioned by Margetts (2011) several years before Cambridge Analytica’s use of Facebook data for political targeting – has led to new modes of misuses and democratic challenges in contemporary society (see Bodó, Helberger, & de Vreese, 2017, for a special issue on political micro-targeting). Corresponding to this, transparency can be used inadvertently or strategically to produce opacity, as stated by de Laat (2018) in relation to algorithmic decision-making, and by Forsbæk and Oxelheim (2015) with regards to organisational audit and accountability. Similarly, a US-focused case study of the data broker industry makes the case that transparency runs up against “insurmountable structural limitations within the political economy” of this

particular industry and that transparency as a policy approach is “subsumed by a discourse of consumer empowerment that has been rendered meaningless in the contemporary environment of pervasive commercial surveillance” (Crain, 2018, p. 89). That is, there seems to be limitations in transparency as a policy response for this type of industry, both at a structural level, as well as a regulatory deflection strategy, at worst only creating “an illusion of reform” (Crain, 2018, p. 89).

In research on AI in computer science the concept of *explainability* (xAI) represents what could be called a “model-close” research strand of relevance to transparency in AI (see Lepri et al., 2018; Ribeiro, Singh, & Guestrin, 2016). XAI is often described as a means to deal with “black box models” (see Biran & Cotton, 2017; Guidotti et al., 2018) or what de Laat (2018) refers to as “inherent opacity” (2018). This xAI-notion of transparency is narrower and more algorithmic model-oriented than for example the necessary transparency (and “explicability”) expressed by AI HLEG (2019) to achieve an ethically sound and trustworthy AI. However, and as noted by Mittelstadt, Russell and Wachter (2019), explanations of machine learning models and predictions can serve many functions and audiences: explanations can be necessary to comply with relevant legislation (Doshi-Velez et al., 2017), verify and improve the functionality of a system, and arguably enhance the trust between individuals, subject to a decision, and the system itself (see Citron & Pasquale, 2014; Hildebrandt & Koops, 2010; Zarsky, 2013).

2. A MULTIDISCIPLINARY NOTION OF TRANSPARENCY

When focusing on transparency in the context of AI, the literature often refers to explainability with reference to both interpretability as well as trust in the systems (see Ribeiro, Singh, & Guestrin, 2016). For example, when assessing users’ trust in applied AI, an assumption made in recent literature (see Miller, 2019) is that the issue of transparency has to take into consideration how ordinary humans understand explanations, and how they assess their relationship to a service, product or company. The development of explainable AI is then, arguably, driven by evidence that many AI applications are not used in practice, partly due to users lacking trust in them (see Linegang et al., 2006). The following hypothesis is then that by building more explainable systems, users would be better equipped to understand and thereby trust the intelligent agents or predictive modelling (Mercado et al., 2016; see also Kopitar, Cilar, Kocbek, & Stiglic, 2019, for a medical example). Trust and its links to transparency, and its required conditions, have been studied in many social-scientific disciplines, including law, over a long period of time. However, research on explainable AI typically does not cite or build on explanatory frameworks based in social science (Miller, Howe, & Sonenberg, 2017; see also de Graaf & Malle, 2017). More could be done with regards to this (see Felzmann, Villaronga, Lutz, & Tamò-Larrieux, 2019, on a “relational” understanding of transparency in AI).

As the opening of the ‘black box’ may bring a number of challenges of a legal, technological and conceptual nature, suggested by Wachter, Mittelstadt and Russell (2017), the notion of transparency in AI – as applied on markets and interacting with humans and institutions – could benefit from a wider definition than the more narrowly defined xAI (see Mittelstadt, Russell, & Wachter, 2019). Drawing from research in law, the social sciences and the humanities, the xAI domain could be complemented with a range of aspects of relevance for AI transparency (argued for in Larsson, 2019), such as:

1. legal aspects of *proprietaryship*, as code and data enters competitive markets (Pasquale, 2015), including trade secrets (Wachter, Mittelstadt, & Russell, 2017); Described by Burrell

- (2016) as an aspect of *intentional opacity* (de Laat, 2018).
2. the need to *avoid abuse*, indicating that too much openness in the wrong context may actually defeat the purpose of an AI-enabled process (Caplan, Donovan, Hanson, & Matthews, 2018; Miller, 2019). There can be incentives for “gaming the system” – examined by de Laat (2018) in terms of “perverse effects of disclosure” – affecting everything from trending topics on Twitter to security issues and welfare distribution.
 3. data and algorithm user *literacy*, indicating that ordinary users’ basic understanding has a direct effect on transparency in applied AI (Burrell, 2016; Haider & Sundin, 2019). This relates to educational efforts in improving literacy, for example on ‘computational thinking’ (Heintz, Mannila, & Färnqvist, 2016);
 4. *the symbols and metaphors* used for communication, that is, mathematically founded algorithms may be dependent on translations to human imaginaries and languages, for example in automated decisions or user agreements (Larsson, 2017; 2019). As concluded by Mittelstadt, Russell and Wachter (2019), there is a fundamental distinction between explainability models and explanations in philosophy and sociology, that is, everyday explanations for humans are contrastive, selective, and social (Miller, 2017), which is not the same as the “interpretability” and explainability found in the xAI domain.
 5. the *complexity* of data ecosystems and markets trading in consumer-data have an unquestionable effect on transparency with regards to the obscure origins of the underlying data and how personal data travels (Christl, 2017; Crain, 2018; Larsson, 2018; Pasquale, 2015); This relates to the debate around the search for a “right to an explanation” in the General Data Protection Regulation (GDPR) – by some feared to lead to a “transparency fallacy” (Edwards & Veale, 2017), and, lastly,
 6. the obscuring effects of *distributed*, personalised outcomes that create challenges not the least for supervisory agencies with limited access and overview attempting to detect structural discrimination or other unfair outcomes (Larsson, 2018; see Larsson et al., 2019).

In this wider notion of transparency in AI, a key challenge from a governance perspective – as AI is applied and interacting with users, consumers and citizens – is arguably to find an appropriate balance between legitimate but not necessarily compatible interests. For example, as noted in the first draft of ethics guidelines from HLEG, there might be “fundamental tensions between different objectives (transparency can open the door to misuse; identifying and correcting bias might contrast with privacy protections)” (AI HLEG, 2018, p. 6). Thus, the interplay between AI technologies and societal values – the applied ethics, social and legal norms – underscores the importance of combining social scientific and contributions from the humanities to computer scientifically based AI research (see Dignum, 2019). This is an argument in line with what Rahwan (2018) has emphasised in terms of a need to keep society “in-the-loop” in order to enable such balances.

3. ETHICAL AND LEGAL RELEVANCE OF TRANSPARENCY IN AI

Transparency in AI has increasingly been highlighted in regulatory development, company policies as well as ethical guidelines over the last few years. For example, the EU adopted a strategy on AI in April 2018, and appointed the High-Level Expert Group (AI HLEG) mentioned above to give advice on both investment strategies as well as ethical issues with regards to AI in Europe. In December 2018, the European Commission presented a coordinated plan – “made in Europe” – prepared with member states to foster the development and use of AI in Europe. By mid-2019, all member states were expected to have their own strategies in place, which was not entirely the case (Van Roy, 2020). The coordinated plan from 2018 included four key areas: increasing investment, making more data available, fostering talent and ensuring trust. The last point, on how to ensure trusted, ethically aligned and trustworthy applications and development

of AI was also in focus for the AI HLEG report published in April 2019. Ethics guidelines as a tool for AI governance is in line with a global trend (Larsson, forthcoming). Jobin et al. (2019) mapped and analysed the current corpus of principles and guidelines on ethical AI. They conclude that of the 84 analysed guidelines, 88% have been published after 2016 and that the most common concept argued for is ‘transparency’. AI HLEG’s guidelines contain an assessment list for practical use by companies that was tested by over 350 organisations during the second half of 2019, and the expert group will finalise a revised version during 2020. According to the European Commission, a key result of the feedback process is that “while a number of the requirements are already reflected in existing legal or regulatory regimes, those regarding transparency, traceability and human oversight are not specifically covered under current legislation in many economic sectors” (2020, p. 9). Another important mode of governance is standardisation, which can be seen in how the IEEE has a working group (P7001) for standardising transparency of autonomous systems, as well as how the international standardisation body ISO does an overview of ethical and societal concerns in AI (ISO/IEC JTC1/SC 42 Artificial intelligence).

Hence, the advocacy for the importance of transparency in AI comes in different forms and is made by different types of stakeholders. While the regulatory field is too rich in relation to transparency in AI to be thoroughly accounted for here, there are at least three important points raised in recent literature that may be mentioned. Firstly, and as pointed out by AI HLEG (2019), important areas are already regulated in the European Union, such as in data protection, privacy, non-discrimination, consumer protection, and product safety and liability rules. Secondly, there are specific provisions that are particularly debated, such as the seeming right for data subjects “to obtain an explanation of the decision reached” where automated processing (GDPR, Art. 22) is involved (preamble 71). For example, Edwards and Veale (2017) state that the law is “restrictive, unclear, or even paradoxical concerning when any explanation-related right can be triggered” (p. 18; see also Felzmann, Villaronga, Lutz & Tamò-Larrieux, 2019; Wachter, Mittelstadt, & Floridi, 2017). Edwards and Veale (2017) argue that even if it was a clear right warranted by the GDPR, the legal conception of explanations as “meaningful information about the logic of processing” may not be provided by the kind of machine learning explanations computer scientists have developed in response (compare to point 4 above). In addition to data protection, there are calls for more studies on how administrative law should adapt to more automated forms of decision-making (e.g., Cobbe, 2019; see also Oswald’s (2018) review of a number of long-standing rules in English administrative law designed to regulate the discretionary power of the state). Thirdly, there are fields addressing transparency in AI that will require legal development, perhaps on ‘algorithmic auditing’ (Casey, Farhangi, & Vogl, 2019) or risk-adapted requirements (see European Commission, 2020; Datenethikkommission, 2019). There are also arguments suggesting that some notions in contemporary data protection, to use an example, might not be well-fitted to current and coming uses of AI and machine learning abilities to gain insights from large amounts of individuals’ data. Hence, Wachter & Mittelstadt (2018) argue for a “right to reasonable inferences”.

CONCLUSION

Transparency in AI plays a very important role in the overall strive to develop more trustworthy AI as applied to markets and in society. It is particularly trust and issues of accountability that drive the contemporary value of the concept, including the narrower scope of transparency found in xAI. At the same time, ‘transparency’ has multiple uses in various disciplines, and

comes with a history from the 1990s. Transparency in AI, or what we interchangeably call AI transparency, takes a system's perspective rather than focusing on the individual algorithms or components used. It is therefore a less ambiguously broad term than algorithmic transparency. In order to understand transparency in AI as an applied concept, it has to be understood in context, mitigated by literacies, information asymmetries, "model-close" explainability as well as a set of competing interests. Transparency in AI, consequently, can best be seen as a balancing of interests and a governance challenge demanding multidisciplinary development to be adequately addressed.

REFERENCES

- AI HLEG, High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI*. The European Commission. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- Ananny, M., & K. Crawford (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Ball, C. (2009). What is transparency? *Public Integrity*, 11(4), 293–308. <https://doi.org/10.2753/PIN1099-9922110400>
- Biran, O., & Cotton, C. (2017) Explanation and justification in machine learning: A survey *IJCAI-17 Workshop on Explainable AI*. http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf
- Bodó, B., Helberger, N., Irion, K., Zuiderveen Borgesius, K., Moller, J., van de Velde, Bol, N., van Es, B., & de Vreese, C. (2018). Tackling the algorithmic control crisis – The technical, legal, and ethical challenges of research into algorithmic agents. *Yale Journal of Law and Technology*, 19(1). <https://digitalcommons.law.yale.edu/yjolt/vol19/iss1/3/>
- Bodó, B., Helberger, N., & de Vreese, C. H. (2017). Political micro-targeting: a Manchurian candidate or just a dark horse? *Internet Policy Review*, 6(4). <https://doi.org/10.14763/2017.4.776>
- Burrell, J. (2016). How the machine thinks: understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). <https://doi.org/10.1177/2053951715622512>
- Caplan, R., Donovan, J., Hanson, L., & Matthews, J. (2018). *Algorithmic Accountability: A Primer*. Data & Society. <https://datasociety.net/library/algorithmic-accountability-a-primer/>
- Casey, B., Farhangi, A., & Vogl, R. (2019) Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise, *Berkeley Technology Law Journal*, 34, 145–189. https://btlj.org/data/articles2019/34_1/04_Casey_Web.pdf
- Christl, W. (2017). *Corporate Surveillance in Everyday Life: How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions*. Cracked Labs. <https://crackedlabs.org/en/corporate-surveillance>
- Christensen, L.T., & Cornelissen, J. (2015). Organizational transparency as myth and metaphor. *European Journal of Social Theory*, 18(2), 132–149. <https://doi.org/10.1177/1368431014555256>
- Citron, D.K., & Pasquale, F. (2014) The scored society: due process for automated predictions. *Washington Law Review*, 89(1). <https://digitalcommons.law.uw.edu/wlr/vol89/iss1/2>
- Cobbe, J. (2019). Administrative law and the machines of government: judicial review of automated public-sector decision-making. *Legal Studies*, 39(4), 636–655. <https://doi.org/10.1017/lst.2019.9>
- Crain, M. (2018). The limits of transparency: Data brokers and commodification. *New Media & Society*, 20(1), 88–104. <https://doi.org/10.1177/1461444816657096>

Datenethikkommission. (2019). *Opinion of the Data Ethics Commission*. Data Ethics Commission, German Federal Ministry of Justice and Consumer Protection. https://www.bmjv.de/DE/Themen/FokusThemen/Datenethikkommission/Datenethikkommission_EN_node.html

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer International Publishing. https://doi.org/10.1007/978-3-030-30371-6_6

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Shieber, S., Waldo, J., Weinberger, D., & Wood, A. (2017). *Accountability of AI under the law: The Role of Explanation*. arXiv. <https://arxiv.org/abs/1711.01134v1>

Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke Law & Technology Review*, 16(1), 18–84. <https://scholarship.law.duke.edu/dltr/vol16/iss1/2>

European Commission. (2020). *White Paper on Artificial Intelligence: a European approach to excellence and trust* (White Paper No. COM(2020) 65 final). European Commission. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Fast, E., & Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. In *Thirty-First AAAI Conference on Artificial Intelligence*. <https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14581>

Fecher, B., & Friesike, S. (2014). Open science: one term, five schools of thought. In S. Bartling, & S. Friesike (Eds.), *Opening science* (pp. 17-47). Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_2

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1), 1–14. <https://doi.org/10.1177/2053951719860542>

Fenster, M. (2015). Transparency in search of a theory. *European Journal of Social Theory*, 18(2), 150–167. <https://doi.org/10.1177/1368431014555257>

Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People — An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707. <https://doi.org/10.1007/s11023-018-9482-5>

Forssbäck, J., & Oxelheim, L. (2015). “The multifaceted concept of transparency.” In Forssbaeck, J., & Oxelheim, L. (eds.). *The Oxford handbook of economic and institutional transparency*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199917693.013.0001>

Fox, J. (2007). The uncertain relationship between transparency and accountability. *Development in Practice*, 17(4-5), 663–671. <https://doi.org/10.1080/09614520701469955>

Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., & Schafer, B., (2018). AI4People — An ethical framework for

a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>

Gasser, U., & Almeida, V. A. (2017). A layered model for AI governance. *IEEE Internet Computing*, 21(6), 58–62. <https://doi.org/10.1109/mic.2017.4180835>

de Graaf, M. M. A., & Malle, B. F. (2017). How People Explain Action (and Autonomous Intelligent Systems Should Too). *2017 AAAI Fall Symposium Series*. AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction. <https://www.aaai.org/ocs/index.php/FSS/FSS17/paper/view/16009>

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5). <https://doi.org/10.1145/3236009>

Haider, J., & Sundin, O. (2018). *Invisible Search and Online Search Engines: The ubiquity of search in everyday life*. Routledge.

Hansen, H. K., Christensen, L. T., & Flyverbom, M. (2015) Introduction: Logics of transparency in late modernity: Paradoxes, mediation and governance. *European Journal of Social Theory* 18 (2), 117-131. <https://doi.org/10.1177/1368431014555254>

Heintz, F., Mannila, L., & Färnqvist, T. (2016) A Review of Models for Introducing Computational Thinking, Computer Science and Computing in K-12 Education. In *Proceedings of the 46th Frontiers in Education (FIE)*. <https://doi.org/10.1109/fie.2016.7757410>

Hildebrandt, M., & Koops, B.-J. (2010). The Challenges of Ambient Law and Legal Protection in the Profiling Era. *The Modern Law Review*, 73(3), 428–460. <https://doi.org/10.1111/j.1468-2230.2010.00806.x>

Hood, C. (2006). Transparency in historical perspective. In C. Hood, & D. Heald (Eds.), *Transparency: The Key to Better Governance?* (pp. 3–23). Oxford University Press. <https://doi.org/10.5871/bacad/9780197263839.003.0001>

IEEE (2019) *Ethically Aligned Design. First Edition. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. <https://ethicsinaction.ieee.org/>

Jakobsson, P. (2012). *Öppenhetsindustrin* [The openness industry] [PhD Thesis]. Örebro University.

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

Kemper, J., & Kolkman, D. (2018). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, 24(14), 2081–2096. <https://doi.org/10.1080/1369118x.2018.1477967>

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). *A governance framework for algorithmic accountability and transparency* (Study No. PE 624.262) Panel for the Future of Science and Technology, Scientific Foresight Unit (STOA), European Parliamentary Research Service.

[https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU\(2019\)624262_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf)

Koivisto, I. (2016). *The anatomy of transparency: the concept and its multifarious implications* (EUI Working Paper No. MWP 2016/09). Max Weber Programme for Postdoctoral Studies, European University Institute. <http://hdl.handle.net/1814/41166>

Kopitar, L., Cilar, L., Kocbek, P., & Stiglic, G. (2019). Local vs. Global Interpretability of Machine Learning Models in Type 2 Diabetes Mellitus Screening. In M. Marcos, J. M. Juarez, R. Lenz, G. J. Nalepa, S. Nowaczyk, M. Peleg, J. Stefanowski, & G. Stiglic (Eds.), *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems* (pp. 108–119). Springer. https://doi.org/10.1007/978-3-030-37446-4_9

de Laat, P. B. (2018). Algorithmic decision-making based on machine learning from Big Data: Can transparency restore accountability? *Philosophy & technology*, 31(4), 525–541. <https://doi.org/10.1007/s13347-017-0293-z>

Larsson, S. (in press). On the Governance of Artificial Intelligence through Ethics Guidelines, *Asian Journal of Law and Society*.

Larsson, S. (2019). The Socio-Legal Relevance of Artificial Intelligence. *Droit et Société*, (103), 573–593. <https://doi.org/10.3917/drs1.103.0573>

Larsson, S. (2018). Algorithmic Governance and the Need for Consumer Empowerment in Data-driven Markets. *Internet Policy Review*, 7(2). <https://doi.org/10.14763/2018.2.791>

Larsson, S. (2017). *Conceptions in the Code. How Metaphors Explain Legal Challenges in Digital Times*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780190650384.001.0001>

Larsson, S. (2014). Justice ‘Under’ Law – The Bodily Incarnation of Legal Conceptions Over Time. *International journal for the Semiotics of Law*, 27(4), 613–626. <https://doi.org/10.1007/s11196-013-9341-x>

Larsson, S., Anneroth, M., Felländer, A., Felländer-Tsai, L., Heintz, F., & Cedering Ångström, R. (2019). *Sustainable AI: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence*. AI Sustainability Center. http://www.aisustainability.org/wp-content/uploads/2019/11/Socio-Legal_relevance_of_AI.pdf

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press. □

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books.

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. In B. Goertzel, & P. Wang (Eds.), *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms. Proceedings of the AGI Workshop 2006* (pp. 17–24). IOS Press.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology*, 31, 611–627. <https://doi.org/10.1007/s13347-017-0279-x>

Linegang, M. P., Stoner, H. A., Patterson, M. J., Seppelt, B. D., Hoffman, J. D., Crittendon, Z. B., & Lee, J. D. (2006). Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(23), 2482–2486.

<https://doi.org/10.1177/154193120605002304>

Margetts, H. (2011). The internet and transparency. *The Political Quarterly*, 82(4), 518–521.

<https://doi.org/10.1111/j.1467-923X.2011.02253.x>

Martinez, R. (2019). Artificial Intelligence: Distinguishing Between Types & Definitions. *Nevada Law Journal*, 19(3), 2015–1042. <https://scholars.law.unlv.edu/nlj/vol19/iss3/9/>

McCorduck, P., & Cfe, C. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press.

Mercado, J.E., Rupp, M.A., Chen, J.Y., Barnes, M.J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human-agent teaming for Multi-UxV management, *Human Factors*, 58(3), 401–415. <https://doi.org/10.1177/0018720815621206>

Miller, T., Howe, P., Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum. Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*. Available at <https://arxiv.org/abs/1712.00547v2>

Miller, T. (2019) Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the conference on fairness, accountability, and transparency - FAT* '19*, 279–288. <https://doi.org/10.1145/3287560.3287574>

Monett, D., Lewis, C. W., & Thórisson, K. R. (2020). Introduction to the JAGI Special Issue “On Defining Artificial Intelligence”—Commentaries and Author’s Response. *Journal of Artificial General Intelligence*, 11(2), 1–100. <https://doi.org/10.2478/jagi-2020-0003>

Oswald, M. (2018). Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128). <https://doi.org/10.1098/rsta.2017.0359>

Pasquale, F. (2015). *The Black Box Society. The Secret Algorithms That Control Money and Information*. Harvard University Press.

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20(1), 5-14. <https://doi.org/10.1007/s10676-017-9430-8>

Reddy, M. (1979) The Conduit Metaphor: A Case of Frame Conflict in our Language about Language. In A. Ortony (Ed.), *Metaphor and Thought* (pp. 284–324). Cambridge University Press. □

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference*

on knowledge discovery and data mining, 1135–1144.

<https://doi.org/10.1145/2939672.2939778>

Ruijter, E., Grimmelikhuijsen, S., & Meijer, A. (2017). Open data for democracy: Developing a theoretical framework for open data use. *Government Information Quarterly*, 34(1), 45–52.

<https://doi.org/10.1016/j.giq.2017.01.001>

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller, A. (2016). *Artificial intelligence and life in 2030* (Study Panel Report 2015-2016). <https://ai100.stanford.edu/2016-report>

Theodorou, A., Wortham, R. H., & Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science*, 29(3), 230–241. <https://doi.org/10.1080/09540091.2017.1310182>

Van Roy, V. (2020) *AI Watch - National strategies on Artificial Intelligence: A European perspective in 2019* (JRC Technical Report No. EUR 30102 EN / JRC119974). Publications Office of the European Union. <https://doi.org/10.2760/602843>

Wachter, S., & Mittelstadt, B. D. (2018). A right to reasonable inferences: re-thinking data protection law in the age of Big Data and AI. *Columbia Business Law Review*, 2019(2).

<https://journals.library.columbia.edu/index.php/CBLR/article/view/3424>

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), 76–99. <https://doi.org/10.1093/idpl/ix005>

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2). <https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf>

Zarsky, T. (2013). Transparent predictions. *University of Illinois Law Review*, 2013(4), 1503–1570. <http://illinoislawreview.org/wp-content/ilr-content/articles/2013/4/Zarsky.pdf>

FOOTNOTES

1. See for example the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), <https://fatconference.org>

2. The bibliometric analysis has been assisted by Fredrik Åström, Associate Professor at Lund University, Sweden.