Three Levels of AI Transparency

Haresamudram, Kashyap; Larsson, Stefan; Heintz, Fredrik

Published in:
Computer

DOI:
10.1109/MC.2022.3213181

2022

Document Version:
Peer reviewed version (aka post-print)

Link to publication

Citation for published version (APA):
Haresamudram, K., Larsson, S., & Heintz, F. (Accepted/In press). Three Levels of AI Transparency. *Computer*. https://doi.org/10.1109/MC.2022.3213181

Total number of authors:
3

Digital Object Identifier: 10.1109/MC.2022.3213181

Three Levels of AI Transparency

Kashyap Haresamudram, *Doctoral Researcher, Lund University, Sweden.* Stefan Larsson, *Associate Professor, Lund University, Sweden.* Fredrik Heintz, *Professor, Linköping University, Sweden.*

Abstract—Transparency is generally cited as a key consideration towards building Trustworthy AI. However, the concept of transparency is fragmented in AI research, often limited to transparency of the algorithm alone. While considerable attempts have been made to expand the scope beyond the algorithm, there has yet to be a holistic approach that includes not only the AI system, but also the user, and society at large. We propose that AI transparency operates on three levels, *(1) Algorithmic Transparency*, *(2) Interaction Transparency*, and *(3) Social Transparency*, all of which need to be considered to build trust in AI. We expand upon these levels using current research directions, and identify research gaps resulting from the conceptual fragmentation of AI transparency highlighted within the context of the three levels.

Index Terms—Artificial Intelligence, Transparency, Explainability, Trustworthy AI, Interaction

1 Introduction

Transparency is often viewed as a prerequisite for trust in society [1]. And in relation to AI, transparency has been highlighted as one of the key ethical considerations required to build trustworthy AI [1]. Particularly in sociology and political science, transparency has been studied extensively and believed to lead to greater trust in groups and institutions [1]. The conversation around transparency in AI has developed relatively recently, rooted in governance of AI and closely related to its conceptual roots in socio-legal discourse. However, it has often been argued that AI transparency is needed in order to build trust in the decision-making of AI systems, and also understand their implication on the larger socio-political and cultural context within which they exist and operate. With AI systems, particularly decision-making and recommender systems, being deployed in all domains from healthcare and law-enforcement to retail and e-commerce, questions regarding whether the algorithm is accurate and should be trusted require opaque 'black-box' algorithms to become transparent [1]. While transparency is generally useful in the case of decision-making systems, especially when decisions are being suggested to aid human decision-makers, it isn't entirely clear whether the same is true for all types of AI systems and contexts of application. Additionally, this is transparency of the algorithms alone, outside of its situated context, and excluding the user interactions. Such a specific definition of transparency, arguably, is unlikely to have an effect on trust. On the other hand some research has found transparency to lead to information overload, and negatively affect trust in consumers [2]. However, in general, there is a need for more empirical research on transparency requirements from a user perspective, in various contexts, for any real conclusions to be drawn [3]. We believe that the current scarcity of such research is the result of a fragmented understanding of AI transparency, and highlight the need to expand the conceptual scope of AI transparency to not only include the AI system, but also the various stakeholders interacting with the system, the context of use of the system, and the larger social implications of its continued use.

While conceptually transparency is rooted in a socio-legal context, within AI research, it has come to be predominantly understood as transparency of the algorithm, closely related to the emerging field of Explainable Artificial Intelligence (XAI) [1]. However, XAI has been criticised for being techno-centric, led by individual XAI researchers' intuition on explanations [3], and with limited consideration to the existing research on both transparency and explanations within the social sciences [4]. This phenomenon has been described as 'inmates running the asylum' [4]. And within this context, transparency is defined simply as the ability to understand an algorithm and its decision making, through nuances such as simulatability, decomposability and algorithmic transparency, all of which focus on the algorithm [5]. We argue that this is a narrow conceptualisation. AI transparency should extend beyond the algorithm into the entire life-cycle of AI development and application, incorporating various stakeholders.

Larsson and Heintz have argued for a broader conceptualisation of AI transparency beyond the algorithm, and elaborated upon the socio-legal context of AI transparency [1][6], although, it still remains domain-specific. But it can be useful to understand how these domain-specific conceptualisations are interconnected, and we argue that a wider framing of the concept of AI transparency can help achieve that. We build that argument within the context of AI by identifying three distinct levels at which AI transparency can be realised, distinguishing three central elements in applied AI, the AI system, the user, and the social context. Bringing transparency across these elements together, we envision a cross-domain framework to build truly transparent, and consequentially, trustworthy AI. We conceptualise these levels as, (1) *algorithmic transparency*, as seen above in XAI, (2) *interaction transparency*, realised through human-AI interaction, and (3) *social transparency*, realised through institutions, laws, and socio-cultural norms. This, we believe, can serve as a road-map to better organise and prioritise gaps in trustworthy AI research, enable a clearer understanding of the larger social context of AI, and help identify cross-domain collaboration opportunities for various stakeholders in AI research and development.

Digital Object Identifier: 10.1109/MC.2022.3213181

1.1 Terminology

This section serves to clarify what we mean by some of the key terms we use in this paper. Of late, AI research has been inundated with numerous, overlapping, sometimes interchangeable terms, describing various associated concepts. The multidisciplinary nature of current conversation around AI also means that the same terms can sometimes be understood differently in different fields. This could potentially be one of the reasons for the fragmented understanding of transparency. To alleviate any misinterpretation, this is our key to the terminology used in this paper.

1.1.1 *Trustworthy AI*

According to the Ethics Guideline for Trustworthy AI outlined by the European Commission High-Level Expert Group on AI, for AI to be trustworthy it must meet three broad criteria, it must be (1) Lawful, (2) Ethical, and (3) Robust [7]. Evidently, this is a very broad definition. The concept of trust itself does not have a universally accepted definition. Transparency is clearly highlighted in the guideline as a crucial element of trustworthy AI.

1.1.2 *AI Transparency*

We use the term AI transparency as an umbrella term encompassing several notions of the concept of transparency from various disciplines that speak to making AI more understandable and human-compatible both individually and societally [1][6]. In this paper we propose sparse usage of this term (and transparency in general) in favour of the three specific levels of AI transparency that better articulate the different contexts and stakeholders involved. This, we suggest, will help alleviate confusion arising from the myriad of understandings of transparency.

1.1.3 *Explainable AI*

Explainable AI or XAI can be defined as algorithms that explicitly consider human comprehensibility of their decisions as a criteria in their computations [8]. They encompass tools and methods to explain algorithmic predictions made by black-box AI. Generally, they tend to produce post-hoc explanations. Currently, the field of Explainable AI deals with a wide variety of research, from highly computational and algorithmic, to methods of representation of information.

1.1.4 *Explanation*

While explainability is the ability to explain algorithmic decision making in human-compatible terms, the explanation itself is a much more qualitative element pertaining to the nature of information exchange between the human and AI (in the context of AI) [8]. Explainable AI deals with the algorithm, but the explanation itself has nothing to do with the algorithm, rather, it speaks more to the resultant interaction between the the AI and the user. We make this distinction explicit, since the criticism of XAI is precisely that while AI developers may have an understanding of XAI methods, in most cases they probably do not have a nuanced understanding of explanations [8].

2 Levels of AI Transparency

Conceptualising AI transparency beyond the algorithm is not a novel endeavour. Several scholars have proposed their own frameworks. Wortham [9] highlights system transparency and organisational transparency as being key to build trust in AI. And Larsson [6] expands upon transparency in the legal context with seven nuances of the concept; proprietorship, avoiding abuse, literacy, data ecosystems, distributed/personalised outcomes, algorithmic transparency, and concepts, terminology and metaphor. While these conceptualisations make significant expansions over the widely used concept of algorithmic transparency, they do not touch upon all aspects of AI development and use. We believe such a holistic approach is needed to truly achieve trustworthy AI.

AI systems are not only algorithms, but through their use, give rise to complex interactions between individuals and devices, within specific contexts and environments, which are in-turn governed by social norms, cultural expectations, and laws. The complex interplay of these interactions is, we have found, not adequately captured in AI transparency research. Conceptually, Meijer [10] distinguishes three broad perspectives on transparency; transparency as a virtue, relation and system. The first perspective encapsulates transparency as norm or inherently desirable value in public actors. The second perspective captures the relational notion where one actor is made transparent to another, and transparency exists as a consequence of this relationship. The third perspective speaks about the complex network of relations that exist within a system that work together to produce transparency [10]. Echoing Meijer's perspectives in the context of AI, we propose an overarching framework where transparency is realised on three levels, the AI system/algorithm, the user interaction, and the social context. They can roughly be seen as representing transparency within the AI system, between the AI and the human user, and between the AI and society at large, see Fig.1. However, while Meijer [10] seems to treat the perspectives as three separate views on transparency, we argue that the three levels we propose in relation to AI are inherently connected, likely even interdependent, and work together to make AI systems transparent.

Broadly, this conceptualisation is in some ways aligned with the ethos of the 'Transparency by Design' framework by Felzmann et al [11], and we view their 9 principles as complementary to our framework. The principles being - "(1) Be proactive, not reactive, (2) Think of transparency as an integrative process, (3) Communicate in an audience-sensitive manner, (4) Explain what data is being used and how it is being processed, (5) Explain decision-making criteria and their justifiability, (6) Explain the risk and risk mitigation measures, (7) Ensure inspectability and auditability, (8) Be responsive to stakeholder queries and concerns, and (9) Report diligently about the system" [11]. However, it can be argued that the principles largely pertain to algorithmic transparency (with some exceptions), and relate to specific set of stakeholders. Through the three levels we seeks to cast a broader lens on transparency. The principles however prove useful in further elaborating upon some concepts covered here.

Digital Object Identifier: 10.1109/MC.2022.3213181

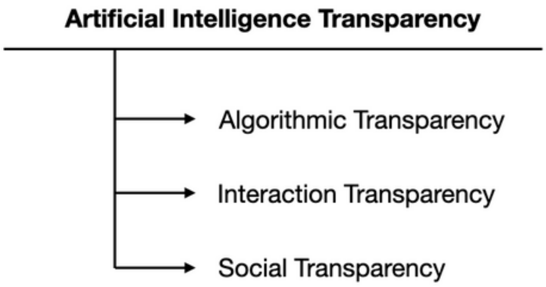

Fig. 1. Levels of AI Transparency

It is important to note that the relevance of each individual level may vary for different stakeholders. For example, algorithmic transparency may be crucial to developers and auditors of the system, but not necessarily as important to the end user, to whom interaction transparency might take precedence. The three levels can be used to create stakeholder maps that can help identify collaboration needs and opportunities when developing and deploying AI algorithms. For example, the hypothetical stakeholder map in Table 1 shows an overview of how the levels of transparency may potentially map on-to various stakeholders (in terms of relevance). Neither the stakeholders, nor the mapping presented here are exhaustive and will likely differ based on the AI system in question and the context it is used in. But we propose that creating such stakeholder maps can help identify areas of cross collaboration as well as explicitly address transparency requirements on different levels.

	Algorithmic	Interaction	Social
Developer	X		
Designer		X	
Owner			X
User		X	X
Regulator	X		X
Society			X

TABLE 1
Example Stakeholder Interest Map

2.1 Algorithmic Transparency

Most widely (mis)understood as AI transparency in general, algorithmic transparency is seemingly the most researched and well understood of the three levels. The primary problem being that complex AI systems process humanly unmanageable amounts of data in humanly incomprehensible ways, resulting in unknown biases in the resulting decisions. Such algorithms are sometimes referred to as ‘black-boxes’. Several notable examples of black-box algorithms exist that have made biased decisions not intended by its developers, for example, Amazon’s AI recruiting tool was found to be biased against women , and Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a criminal recidivism prediction algorithm employed by some states in the US, that has been shown to be racially biased against African-Americans [6].

Closely associated is the concept of openness. The ability to access and scrutinise code, data sets, and accompanying systems is essential for accountability, and an important part of AI transparency [1]. The two examples listed above are proprietary algorithms, making them harder to study and evaluate. Additionally, complex arguments regarding data collection, privacy, biased data, historical data and the like have stemmed from algorithmic transparency research. And there are still important questions to be answered here such as transparency with regards to synthetic data. Algorithmic decisions are generally going to be influenced by the characteristics of the data used to train the algorithm, as well as the data the algorithm is used on. Understanding this interaction between the algorithm and the data is at the core of algorithmic transparency.

Various methods to open these black boxes and ‘explain’ the decisions made by these algorithms have been proposed under a relatively new research domain called Explainable AI. Using several methods, most often secondary AI algorithms designed to decode the primary ‘black-box’ algorithms, the decisions are broken down into human-comprehensible terms. For example, Shapley Additive Explanations (SHAP) is a popular XAI algorithm that produces explanations by highlighting the feature weights within the black-box AI that most influenced a given prediction/decision. This information is presented through graphs [5]. XAI methods however are not always accurate and are also generally inaccessible to non domain experts. The explanations, which often take the form of probabilities or graphical representations can also be unintuitive and convoluted. And also mentioned above, XAI has been criticised for relying on individual XAI researchers’ intuition on explanations [3], and with limited understanding of explanations within the social sciences [4]. We argue that the ’explanation’ itself is less suited as part of algorithmic transparency, and more suited as part of interaction transparency that we elaborate in the next section. Alternatively, other methods for algorithmic transparency have also been proposed without needing post-hoc explanations of algorithmic decisions. Interpretable AI is one such method where simple, human-comprehensible algorithms are used instead of black-box models. It is usually understood that complex models with large data sets yield better and more accurate predictions, however Rudin [12] has recently demonstrated that leaner, inherently interpretable models can be just as effective in certain domains.

Generally though, the solutions to achieve algorithmic transparency tend to cater to domain-experts. This is likely due to a justified bias in research pertaining to ’high-stakes’ AI systems such as in healthcare, where AI is generally used by experts as decision-support systems. Whether algorithmic transparency is needed in everyday contexts to end users, such as during online shopping, is not well understood. Although the limited research that does exist finds that this type of transparency may not matter to users in everyday contexts [13]. Algorithmic transparency is probably most relevant to domain-experts and auditors/regulators.

Digital Object Identifier: 10.1109/MC.2022.3213181

2.2 Interaction Transparency

Miller claims about explainability that, "Ultimately, it is a human–agent interaction problem. Human-agent interaction can be defined as the intersection of artificial intelligence, social science, and human–computer interaction (HCI)" [8]. As we have stated before, we argue instead that it is not explainability, rather the explanation itself that is the human-agent interaction part of the problem (explainability is the algorithmic operationalisation of an explanation).

As AI systems get more complex and advanced, so too does their ability to interact with the users as well as influence their shared environment. The ability of AI systems to learn and adapt to their users brings forth an entirely different interaction paradigm and affordances towards transparency through the interaction. But discourse on AI transparency has evolved as though various elements of interaction don't influence it. Research on how AI transparency translates in an applied setting is limited [13], what it means to the user is not well understood, and how to design for it is not clear either [13]. How transparency translates in interaction is seemingly the least studied of all three levels we have highlighted in this paper.

Tangibility, embodiment and entanglement form a compelling basis for interaction transparency. Increasingly, our interactions with AI are embodied. We wear AI in smart watches, we let AI change and adjust our environment in smart homes, smartphones are intrinsically linked to the social fabric of modern lives, and we experience these devices as an extended continuum of our body. As objects that can be touched and interacted with, the affordances through the materiality of these objects could be used to further embody the experience and entangle the individual with the device. Ghajargar et al [14] conducted a design workshop to ideate and build upon the concept of 'Graspable AI' as an extension of the scope of explainable AI, using tangible and embodied interaction and the material body of the object to create rich, contextual, situated explanations that could enable transparency.

Lakoff and Johnson [15] write extensively about the nature of language and the relational metaphors we use with regards to our body to make sense of the world, making our experiences necessarily *embodied*. The use of metaphors can be extended outside of language to objects, designing interaction possibilities with one object through the embodied experience of another analogous object. For example, representing e-books as real books digitally, and transferring existing knowledge about interaction possibilities with real books onto e-books. "The stronger this coupling, the more natural and pervasive the metaphor(s) involved, the more naturalistic and transparent the interaction becomes" [16]. 'Third-wave HCI' embraces the concepts of tangibility and embodiment to understand knowledge production in interaction. Frauenberger [17] proposes *Entanglement* as a new paradigm in interaction design arguing for objects and the environment as forming equal social actors to human actors within interaction, and that knowledge is co-created between all actors as part of the interaction rather than existing entirely in an objective reality or as an external social phenomenon. This theory aligns perfectly with the argument we make about transparency (knowledge) arising as a result of interaction. To illustrate this, Frauenberger gives the example of a hypothetical device *Flow* that provides information about the ease or anxiety levels of different actors in an interaction, and postulates that the input form the device may become an inherent part of interaction with time, making this new sense (input) shape future interactions, exhibiting an entanglement between the technology and the users [17]. In the context of this paper, this example can be interpreted as enabling a form of transparency, providing information about the state of anxiety of an actor in an interaction, information that would otherwise not be available to the actors. With AI, such entanglements can be used as tools to open new avenues of interactions, as well as transparency.

Ultimately, it is through the interaction that knowledge exchange between the AI system and its users takes place. An intimate coupling of behaviours between AI and user is a form of transparency that could potentially enable a nuanced understanding of the strengths and limitations of an AI system, strengthening human trust in it. Embodied/entangled interaction design can play a much larger role in enabling transparency than has seemingly been recognised, and much research is needed in this space.

2.3 Social Transparency

Today AI, specifically machine learning, is being applied in almost every domain that has access to big data, and domains that do not are rushing to create those opportunities [1]. Big tech has made AI-enabled services ubiquitous, leaving experts and researchers to play catch-up with relevant legal frameworks, and understand its larger social impact.

2.3.1 Law and regulation

With high-profile cases such as Cambridge Analytica still in our recent collective memory, the conversation around data privacy and algorithmic responsibility is highly relevant [6]. Given that applied AI is so commonplace in society, several governments have begun to formulate frameworks to regulate it. Transparency in AI is widely considered an ethical obligation [18]. This statement is evidenced by recent developments within the EU, where a high-level expert group has drafted 7 key facets to evaluate when implementing AI in the Ethics Guideline for Trustworthy AI, highlighting transparency as a key facet [7]. In a recent study, it was found that transparency in some form was mentioned as a the most common element in 84 different ethics guidelines on AI across the world [18].

The approach to AI transparency is necessarily domain and application specific [1]. And this is reflected in how the EU regulates AI as outlined in the proposal for an Artificial Intelligence Act presented in 2021, by dividing the applications into four broad, risk-based buckets, unacceptable risk, high risk, limited risk, and minimal risk [19]. High-risk AI is scrutinised much more heavily, and the transparency requirements of high-risk AI are much higher [19], and operate on multiple levels of transparency as categorised above. Limited-risk AI also has transparency obligations according to the proposed AI regulation, insofar as the users have the right to know when interacting with an AI [19]. This risk-based approach is also echoed in recent

Digital Object Identifier: 10.1109/MC.2022.3213181

work by Rudin [12], who proposes the use of interpretable models as an alternative to black-box AI in high-stakes AI applications. Rudin's [12] work echos a common theme in a majority of AI research, not just in computer science but also the social sciences, with a key emphasis on high stakes AI. Arguably, the risk-based approach is the first attempt in trying to incorporate the situatedness (highly context-dependent nature) of AI systems within a legal framework. However, risk levels are one way to define context in which AI operates, while some form of categorisation is necessary to differentiate between various AI systems and their contexts of use, it remains to be seen whether this is the ideal approach.

2.3.2 Society and Culture

In European consumer and data protection, much emphasis is placed on information as a means of transparency and on individual responsibility towards that information. This has resulted in implementing solutions like cookie consent banners for transparency in data collection. However, these individual privacy agreements are far too many, causing information overload; indicating a flawed approach [2]. Critics have argued that a collective approach is likely more beneficial to society, and have advocated for institutionalised solutions instead, akin to the institutionalised solutions seen in the aviation or food industries, whereby we as consumers don't need to inspect and build trust in individual companies or products making up the industry. Rather, trust is formed in the system as a whole, whose individual parts are highly regulated by laws that encourage transparency [20]. Closely related to the idea of institutionalised trust is organisational transparency, where transparency enables accountability, thereby forming trust. It isn't known whether such an industry-wide standardisation is possible with regards to AI, but there are indications that consumers would prefer such as solution too.

Overarching the conversation around AI transparency, data privacy and 'datafied' living in general is literacy. Digital literacy is at the core of end users' ability to comprehend the technology they are interacting with, and consequently for transparency to be realised [6]. Studies have found that a majority of users have very little understanding of online data collection [2]. With data being an integral part of AI, and consumer oriented AI relying on online data collection, one can then extrapolate, perhaps, that literacy regarding AI in general is probably extremely low. While some have argued that trustworthy AI should not default to placing the burden of literacy on the consumer, preferring institutions and regulations to mediate instead [20], it is still worrying that active measures in improving literacy are scarce.

Lastly, ethics and norms are not necessarily universally consistent. Larsson writes that, "this could for example regard different groups, ethnicities, religions, demographics with different notions of what is regarded as right and wrong for everything from families, nudity, gender, sexuality, to free speech, media habits, driving behaviour, and so on" [6]. These nuanced, often sensitive, social challenges with regards to AI transparency will require careful consideration.

3 Conclusion

Given that AI transparency is often understood as a prerequisite for trustworthy AI, but at the same time is a fragmented concept, this broad framework expands the scope to include various levels that AI transparency encompasses. The framework provides the ability to identify and weigh different notions of transparency based on the context to enable informed prioritisation. Based on our review, we identify potential research areas that can contribute to the current understating of AI transparency and its role towards trust in AI. Firstly, user-centred research on AI transparency is limited. Much remains to be learned about the user needs with regards to AI transparency in various contexts, as well as the role interaction plays in generating transparency. Secondly, further research needs to be conducted on alternative methods to achieve transparency that don't involve great volumes of information and individual responsibility. More work is needed towards establishing the collective responsibility (institutionalised trust) argument which is seemingly at odds with parts of the current direction of AI regulation worldwide. Lastly, novel approaches in the form of embodied interaction should be embraced and researched to solve novel interaction problems posed by novel technology within the broad AI domain.

In conclusion, given the widely accepted notion that AI transparency can greatly contribute towards building trustworthy AI, our proposed three-layer approach to AI transparency through (1) Algorithmic transparency, (2) Interaction transparency, and (3) Social transparency, sheds come light on the various stakeholders and contexts involved. It expands the scope of AI transparency beyond the algorithm. And most importantly, it illustrates the complex and multifaceted nature of transparency, and emphasises the need for multidisciplinary research and cross-domain collaboration in the field.

Acknowledgments

This work was funded as part of 'AI Transparency and Consumer Trust' project, by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS).

References

[1] Larsson, S., and Heintz, F. *Transparency in artificial intelligence*, Internet Policy Review, 9(2), 2020.

[2] Larsson, S., Jensen-Urstad, A., and Heintz, F. *Notified But Unaware: Third-Party Tracking Online*, Critical Analysis of Law, 8(1), 2021, 101-120.

[3] Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., and Weisz, J. D. *Expanding explainability: towards social transparency in AI systems*, In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-19), 2021.

[4] Miller, T., Howe, P., and Sonenberg, L. *Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences.*, arXiv preprint arXiv:1712.00547, 2017

[5] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., and Herrera, F. *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*, Information fusion, 58, 2020, 82-115.

[6] Larsson, S. *The socio-legal relevance of artificial intelligence*, Droit et societe, (3), 2019, 573-593.

[7] High Level Expert Group on Artificial Intelligence. *Ethics Guideline for Trustworthy AI*, European Commission, 2019.

Digital Object Identifier: 10.1109/MC.2022.3213181

[8] Miller, T. *Explanation in artificial intelligence: Insights from the social sciences*, Artificial intelligence, 267, 2019, 1-38.

[9] Wortham, R. H. *Transparency for Robots and Autonomous Systems: Fundamentals, technologies and applications*, Institution of Engineering and Technology, 2020

[10] Meijer, A. *Transparency*, The Oxford handbook of public accountability, 2014

[11] Felzmann, H., Fosch-Villaronga, E., Lutz, C., and Tamò-Larrieux, A. *Towards transparency by design for artificial intelligence*, Science and Engineering Ethics, 26(6), 2020, 3333-3361

[12] Rudin, C. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, Nature Machine Intelligence, 1(5), 2019, 206-215.

[13] Felzmann, H., Villaronga, E. F., Lutz, C., and Tamò-Larrieux, A. *Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns*, Big Data and Society, 6(1), 2019.

[14] Ghajargar, M., Bardzell, J., Smith-Renner, A. M., Höök, K., and Krogh, P. G. *Graspable AI: Physical Forms as Explanation Modality for Explainable AI*, In Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction (pp. 1-4), 2022.

[15] Lakoff, George, and Johnson, Mark. *Metaphors we live by*, University of Chicago press, 2008.

[16] Fishkin, Kenneth P., Thomas P. Moran, and Beverly L. Harrison. *Embodied user interfaces: Towards invisible user interfaces*, IFIP International Conference on Engineering for Human-Computer Interaction. Springer, Boston, MA, 1998.

[17] Frauenberger, Christopher. *Entanglement HCI the next wave?*, ACM Transactions on Computer-Human Interaction (TOCHI) 27.1 2019, 1-27.

[18] Jobin, Anna, Marcello Ienca, and Effy Vayena. *The global landscape of AI ethics guidelines*, Nature Machine Intelligence 1.9, 2019, 389-399.

[19] European Commission *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, https://artificialintelligenceact.eu/the-act/,COM(2021) 206 final, 2021.

[20] Knowles, Bran, and John T. Richards. *The sanction of authority: Promoting public trust in AI*, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021.

Kashyap Haresamudram is a Doctoral Researcher at the Department of Technology and Society, Lund University, Sweden. His research interests include AI Transparency, Human-AI Trust, and Human-AI Interaction. Haresamudram received his Master of Arts in Cognitive Semiotics from Aarhus University, Denmark. He is a candidate at The Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS). Contact him at kashyap.haresamudram@lth.lu.se.

Stefan Larsson is a Senior Lecturer and Associate Professor at the Department of Technology and Society, Lund University, Sweden. His research interests include Trust, Transparency, and the socio-legal impact of autonomous and AI-driven technologies. Larsson holds two Doctor of Philosophy degrees in Sociology of Law and Spatial Planning, respectively, as well as a Master of Laws (L.L.M.). Contact him at stefan.larsson@lth.lu.se.

Fredrik Heintz is a Professor at the Department of Computer and Information Science, Linköping University, Sweden. His research interests include AI, Trustworthy AI, and the combination of reasoning and learning. Heintz received his Doctor of Philosophy in Computer Science from Linköping University. He is a fellow of the Royal Swedish Academy of Engineering Sciences. Contact him at fredrik.heintz@liu.se.