Being Transparent about Transparency in the Context of Artificial Intelligence

Clàudia Figueras, Harko Verhagen, and Teresa Cerratto Pargman

Department of Computer and Systems Sciences, Stockholm University, Sweden
claudia, harko, tessy@dsv.su.se
https://dsv.su.se/

Keywords: Artificial Intelligence · Transparency · Sociocognitive technical systems

1 Introduction

What distinguishes AI from other technologies is the use of black-box functions in some of its algorithms, which makes it too complex for humans to fully comprehend its decision making criteria or rationale [12]. Data-driven AI is not open to inspection by a human or the artefact itself as it builds upon deep learning technologies. Furthermore, since AI potentially has an agency level not obtained in technology in general, AI can play a different role than any other technology that can be part of sociotechnical systems. Thus, within the multi-agent systems research community, the term sociocognitive technical systems (SCTS) [3,13] has been proposed to express that AI's agency goes beyond "traditional" technologies. More in particular, AI artefacts in a shared social space with human agents need to "'understand' and reproduce features of the human social mind like commitments, norms, mind reading, power, trust, 'institutional effects' and social macro-phenomena" [3]. Thus, humanities and social and behavioural sciences are all needed to strengthen the sociotechnical systems approach that fits more traditional technology so we can develop and analyse SCTS.

The first step in our project is to approach transparency from different angles. Following this, we analyse an example with respect to the various transparency aspects to show the benefits of these. Finally, we propose the next step for our project.

2 Transparency

Transparency has been adopted as a core value and central standard for several professionals, such as journalism [5], as a hope to rectify the loss in credibility and trust by the general public [9]. With the advent of the use of software in many realms of our lives, transparency has become the gold standard towards achieving accountability [19]. Transparency as an endeavour has some limitations, in any case. First of all, human understanding is commonly linked to a metaphor of "seeing" which, in turn, follows other similar conceptions such as light, clarity

and transparency [11]. This usually relates transparency as a positive pursuit. Moreover, there is the assumption that observing more facts of a phenomenon increases the control we can have over it and the chance to find responsible actors [1,4]. However, high levels of transparency can also be used as means to distract, since critical pieces of information may become inadvertently hidden in the deluge of visible information, what Stohl et al. [18] called *inadvertent opacity*, or if done intentionally, *strategic opacity*. Thus, transparency by itself does not solve any issues; transparency needs to be accompanied by a series of practices. Finally, 'absolute' transparency may threaten privacy and expose vulnerable individuals [1,17], so it must be sought with caution.

3 Example analysis: COMPAS

An example of a widely discussed AI tool is COMPAS, developed by the for-profit company *equivant*. COMPAS predicts the likelihood of a defendant becoming a recidivist using a recidivism risk scale, which is based on the defendant's characteristics and past criminal records in the USA [7,10]. COMPAS uses machine learning algorithms to rank defendants into low, medium or high risk of committing a misdemeanour or felony within two years of assessment for each category [2]. Both the COMPAS algorithm and the data it uses received significant criticism due to relying on factors that could correlate with socioeconomic status, such as the questions "How hard is it for you to find a job ABOVE minimum wage compared to others?" and "Do some of your friends or family feel they must carry a weapon to protect themselves?" [15], which according to [10] biases the resulting prediction of recidivism risk. This possible bias combined with a study that prompted debate on COMPAS unreliability [2] reduced the trust and credibility of the use of AI for decision-making in public institutions [16].

Given the current tendency of incorporating AI systems in public systems, such as in criminal justice as it is the case for COMPAS, a higher demand for transparency has been argued for [5,14], with emphasis on including transparency as a means to ensure that AI systems behave in ways that guarantee human well-being [6]. We must, however, consider which type of transparency we are referring to. Focusing on AI, do we mean transparency of data, algorithms or agency (e.g. AI artefacts pretending to be humans and spreading false information via social media to influence humans' voting preferences)? Defining this is vital to be able to tackle the issue of transparency. We should also note how differently these kinds of transparency affect individuals and society. All in all, the aim of this notion of transparency is not to tick a box once the AI system is developed (i.e. to check that transparency is fulfilled) but to consider it throughout the whole process, from design to deployment. Following the definition of *algorithmic transparency* by Diakopoulous and Koliska ("the disclosure of information about algorithms to enable monitoring, checking, criticism, or intervention by interested parties") [5], we modified and extended it to data and agency transparencies. *Data transparency* is the disclosure of information about the data fed into the algorithms, such as content information as well as the con-

text of the data collection. This could include the purpose of why the data was collected or created, who created the dataset and on behalf of what entity, and who funded the dataset's creation and collection (we refer to Gebru et al. [8] for a more detailed list). *Agency transparency* is the disclosure of information about the agent who is interacting with a human (or with a group of humans), i.e. if we are we dealing with a human agent or an AI artefact.

What is significant here is that while the data COMPAS uses can be found in several public criminal records of the USA, the details of the algorithm have not been revealed by the company that developed it, only part of its results (i.e. machine learning evaluation measures) is divulged, and particularly questions about the agency transparency remain unanswered. Notably, we wish to pose questions about these type of AI systems such as: How aware are the affected people that an AI decision-making system is deciding about them? Can the affected people choose whether an AI or a human is making these decisions? Is it publicly known which institutions apply these AI systems? These questions bring us to engage with the notion of transparency of AI artefacts with the objective to unpack the types of transparency we are referring to and search for a better understanding of the opportunities and challenges of transparency in such AI systems.

4 Future research

Building on Virginia Dignum's principles of Ethics *in, by* and *for* Design [6], we propose Transparency *in, by* and *for* AI Design. Transparency *in* Design points to ensure transparency during the developmental and evaluation processes of AI artefacts. Transparency *by* Design refers to integrating the transparency in the behaviour of AI systems and, if possible, enable the system to reflect about its own transparency. Transparency *for* Design considers the societal impact of the design choices and takes the necessary steps to minimise the harmful effects of a lack or excess of transparency. In the next step we will develop these concepts more thoroughly and discuss these with AI researchers and developers.

References

1. Ananny, M., Crawford, K.: Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society **20**(3), 973–989 (Mar 2018)
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine Bias (May 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Castelfranchi, C.: InMind and OutMind Societal Order Cognition and Self-Organization: The role of MAS. In: Invited talk for the IFAAMAS "Influential Paper Award". Saint Paul, MN, USA (2013), https://www.slideshare.net/sleeplessgreenideas/castelfranchi-aamas13-v2?ref=httpMay2013
4. David, M.: The Correspondence Theory of Truth (May 2015), https://plato.stanford.edu/entries/truth-correspondence/

5. Diakopoulos, N., Koliska, M.: Algorithmic Transparency in the News Media. Digital Journalism **5**(7), 809–828 (Aug 2017)
6. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Artificial Intelligence: Foundations, Theory, and Algorithms, Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30371-6
7. Dressel, J., Farid, H.: The accuracy, fairness, and limits of predicting recidivism. Science Advances **4**(1), eaao5580 (Jan 2018)
8. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for Datasets. arXiv:1803.09010 (Mar 2018)
9. Hayes, A.S., Singer, J.B., Ceppos, J.: Shifting Roles, Enduring Values: The Credible Journalist in a Digital Age. Journal of Mass Media Ethics **22**(4), 262–279 (Oct 2007)
10. Kirkpatrick, K.: It's not the algorithm, it's the data. Communications of the ACM **60**(2), 21–23 (Jan 2017)
11. Larsson, S., Heintz, F.: Transparency in artificial intelligence. Internet Policy Review **9**(2) (May 2020), `https://policyreview.info/concepts/transparency-artificial-intelligence`
12. Mittelstadt, B., Russell, C., Wachter, S.: Explaining Explanations in AI. Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19 pp. 279–288 (2019), arXiv: 1811.01439
13. Noriega, P., Padget, J., Verhagen, H., d'Inverno, M.: Towards a framework requirements for socio-cognitive technical systems. In: Coordination, Organizations, Institutions, and Norms in Agent Systems X, Lecture Notes in Computer Science, vol. 9372. Springer International Publishing, Switzerland (2015)
14. Pasquale, F.: The black box society. Harvard University Press (2015)
15. Northpointe Institute for Public Management, I.: Measurement and Treatment Implications of COMPAS Core Scales. Tech. rep. (2009), `https://www.michigan.gov/documents/corrections/Timothy_Brenne_Ph.D._Meaning_and_Treatment_Implications_of_COMPAS_Reentry_Scales_297503_7.pdf`
16. Rudin, C., Ustun, B.: Optimized Scoring Systems: Towards Trust in Machine Learning for Healthcare and Criminal Justice. Interfaces **48**(5), 449–466 (2018)
17. Schudson, M.: The rise of the right to know: politics and the culture of transparency, 1945-1975. The Belknap Press of Harvard University Press, Cambridge, Massachusetts (2015)
18. Stohl, C., Stohl, M., Leonardi, P.M.: Managing Opacity: Information Visibility and the Paradox of Transparency in the Digital Age. International Journal of Communication **10**, 123–137 (2016)
19. Ward, S.J.: The magical concept of transparency. In: Zion, L., Craig, D. (eds.) Ethics for Digital Journalists: Emerging Best Practices. Routledge, Taylor & Francis Group, New York (2015)