"AI: Coping with Technical Transparency"

International conference organized by the Ethics & AI Chair

of the MIAI at the Université de Grenoble-Alpes

8th-9th October 2020

Presentation of the Chair

The Ethics & AI Chair is part of the Multidisciplinary Institute in Artificial Intelligence (MIAI)[3] and is affiliated to the Institut de Philosophie de Grenoble (IPhiG)[4]. It aims to develop, over a four-year period (2019-2023), a philosophical understanding of artificial intelligence through a sustained dialogue with computer science and robotics, cognitive, social and clinical psychology, information and communication studies, as well as management studies. At the crossroads between political philosophy, public ethics and philosophy of technology, the chair seeks to explore the social, moral and political dimensions at stake in the deployment of AI technologies, in a way that is both critical and attentive to their technical realities.

Conference argument

The large-scale deployment of machine learning processes which actively take part in a wide range of social practices, (cancer screening, credit scoring, cultural content recommendation, etc.) begs the question of how "transparent" they should be. Although this term has been almost universally promoted as an essential democratic value over the past twenty years, its consistency remains difficult to grasp. In fact, it would seem we are confronted with a kind of double transparency, characterized by a tension which is barely questioned or problematized.

On the one hand, machine learning systems with which we interact are still, by and large, developed according to an ideal of transparency that can be qualified as *phenomenological* (M. Wheeler; D. Ihde; M. Heidegger). In this perspective, a properly functioning technical object, one which fulfills its function, disappears in action and fades from the immediate field of perception by allowing amplified possibilities for acting. We are not, for example, aware of the pencil when we use it to write normally but only when its lead breaks; likewise, eyeglasses are only perceptible when they bear an anomaly that disrupts the wearer's normal vision. This logic is commonly transposed on the level of algorithmic systems' design in a *user-friendly* or *plug-and-play* approach. The technical object is thus expected to be as discrete or even invisible as possible, the prevalent idea being that the user should be able *to see through it* as it acts with it. In a related way, this same logic is what underpins the efficacy of algorithms like Google's PageRank, where content publishers are explicitly asked to not take the algorithms parameters into account and to behave "naturally" (D. Cardon). Elsewhere, this logic is at work on platforms such as YouTube, Amazon or Netflix whose recommendations are designed to be as fluid and imperceptible as possible (T. Reigeluth) The question then is whether the normative standard implied by the phenomenological transparency paradigm is transferable or even desirable when it comes to technical objects, which are not simple

[3] https://miai.univ-grenoble-alpes.fr/

[4] https://iphig.univ-grenoble-alpes.fr/programmes-recherche/chaire-ethique-ia

instruments for our senses and organs, but are systems exhibiting a degree of behavioral autonomy (M. Wheeler) and normative inventiveness, that is that they transform the social practices within which they are deployed (J. Grosman and T. Reigeluth). In other words, one of the stakes might actually be to make such technical systems as perceptible as possible or at least to rethink how they present themselves to us in action.

On the other hand, critics of such systems – generally social scientists or legal scholars – demand that the normative effects produced by these systems (biases, discriminations, etc.) be made legible and transparent (Diakopoulos; Pasquale) in a form of governance by accountability. The question, however, as to whether such a criterion is enough to govern the complexity of such technical systems or whether it allows us to access a deeper truth than the one which plays itself out through its normative effects, remains wide open. The paradox of these « social » critiques is that they reproduce a widely held view by the GOFAI stance of cognition in which mental states (in the brain) are transparent and correspond to observable behaviors. Yet, if the engineers who took part in developing machine learning algorithms cannot account for the precise reasons as to why a system produced such or such output deemed to be problematic on a social or ethical level, it begs the question as to whether "opening the black box" will be of any use to their users. And even if it does prove to be useful or even necessary, nothing says that it is *enough* to govern these systems (J. Burrell). Indeed, as a technique of governance, transparency poses a series of difficulties and limits that need to be taken into account: information overload, rule of expertise, marginalization of political stakes, difficulty in determining the concerned public, over-responsibilizing individual resources and competences needed for decoding information (M. Ananny et K. Crawford ; T. Berns).

The risk then is that transparency ends up being a rather weak or loose-ended requirement, with which nobody would consider disagreeing. Who doesn't want transparency when the term has become synonymous with good governance or even democracy? And how do we arrive at a robust political norm that is technically actionable and applicable? Thus, the question we are forced to ask is the following: ***what must be done with transparency?*** This conference is an attempt to answer this question by way of the different fields and approaches concerned by the question of transparency in machine learning systems: cognitive science, philosophy of science and technology, political philosophy, sociology, anthropology, information and communication sciences, management, legal studies, etc. The conference languages are French and English, but it asked that English-speakers have at least a passive understanding of French.

The conference will unfold along three thematic axes, which do not necessarily involve any alignment with disciplinary distinctions:

4. **Technical norms:** How does transparency translate onto a technical level? What are the techniques available for making the system transparent for the user? What does transparency look like in terms of engineering practices? What are the aspects of an algorithmic system which are most important to audit?
5. **Epistemological frameworks:** How can we determine whether a technical system is effectively known? What criteria do we have to make systems legible? What publics and competencies are supposed by technical transparency? What are the contextual or structural limits which hamper full access to the systems functioning?
6. **Political perspectives:** What normative demands can we have towards the design and regulation of machine learning systems? How should these demands be informed by technical norms and epistemological frameworks? Do these systems require specific modes of deliberation and regulation?

General bibliographical references

Wheeler, M. “The reappearing tool: transparency, smart technology, and the extended mind.” *AI & Soc* 34**,** 857–866.

Ihde, Don, *Embodied Technics,* Automatic Press, United States, 2010.

Heidegger, Martin, *Etre et temps,* Editions Gallimard, Paris, 1986 [1927].

Grosman, Jérémy, & Reigeluth, Tyler, “Perspectives on algorithmic normativities: engineers, objects, activities.” *Big Data & Society*, 2019.

Ananny, Mike, & Crawford, Kate, “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability.” *New Media & Society*, *20*(3), 2018.

Berns Thomas, *Gouverner sans gouverner. Une archéologie politique de la statistique.* Presses Universitaires de France, « Travaux pratiques », 2009.

Cardon Dominique, « Dans l'esprit du PageRank. Une enquête sur l'algorithme de Google », *Réseaux*, 2013/1 (n° 177), p. 63-95.

Pasquale, Frank, *The Black Box Society,* Havard University Press, Cambridge, MA, 2015.

Diakopoulos, Nicholas, « Algorithmic Accountability Reporting : On the Investigation of Black Boxes », Tow Center for Digital Journalism*, Columbia School of Journalism*, 2013.

Burrell, Jenna, « How the machine ‘thinks’: Understanding opacity in machine learning algorithms » in *Big Data & Society,* vol.3 n°1, 2016.

Sandvig, Christian, « When the Algorithm Itself Is a Racist: Diagnosing Ethical Harm in the Basic Components of Software », *International Journal of Communication,* vol. 10, 2016, pp. 4972-4990.

-