Meeting Societal Expectations: Cooperative Responsibility in the Design and Governance of AI

Prof. Jo Pierson, Vrije Universiteit Brussel, Belgium (jo.pierson@vub.be)
Dr. Aphra Kerr, Maynooth University, Ireland (aphra.kerr@mu.ie – corresponding author)
&
Dr. Rob Heyman, Vrije Universiteit Brussel, Belgium (rob.heyman@vub.be)

As media are moving towards the ubiquitous capturing and processing of (personal) data, a key challenge is how to maintain trust and counteract manipulation (Susser, Roessler & Nissenbaum, 2019). Despite societal expectations in Europe that we can design ethical AI, and that developers and governments should share responsibility for the outcomes of AI use, it is unclear how we can effectively achieve these goals (Kerr, Barry & Kelleher, 2020). Much effort to date has been on technological solutions to fairness and explainability, and on the potential role of ethics and privacy by design. By contrast we focus on the challenge of responsibility as a potential source of European leadership and competitiveness. We are collaborating on an agenda which takes a multi-level distributed approach to the design and governance of AI. In this presentation we will outline our research agenda and give examples from ongoing and future projects.

Our agenda is to help address the social science deficit in AI research (Sloane & Moss, 2019). We propose a socio-technical perspective which gives equal weight to social and technical innovations. This perspective goes beyond the research focus on optimality, and the business focus on commercial returns, to ask how can we innovate while taking into account European democratic values (like freedom of speech, non-discrimination and data privacy) and the cooperative governance of AI in the public interest. Finally, it considers the domain specificity of deployment, and the requirement for accountability and responsibility to operate throughout the lifecycle. Thus we argue that AI innovations in the media and entertainment industries need to take into account the specificities of these industries in relation to their role in democratic processes, and as cultural practices for diverse audiences.

Our approach suggests that ubiquitous media innovations need to start from a **participatory governance perspective of "cooperative responsibility"** (Helberger, Pierson & Poell, 2018), where the values, norms and expectations of different stakeholders (media sector, researchers, developers, public agencies, citizens, civil society and industry) are clearly identified as early as possible, where responsibilities are acknowledged, and where this knowledge is translated into practical approaches and resources that can be deployed 'on the ground' both in the early phases of technology development (as proposed in designing-by debate (Ausloos et al., 2018)) and post deployment where necessary. This requires a multi-level socio-technical approach which is participatory, co-creative and involves cooperative forms of governance.

What do we mean by this? The participatory co-creative approach starts from co-creative methods (e.g. based on scenario building techniques, futuristic images and theatre approaches) and social interaction with automated decision-making systems in particular contexts, investigated by way of qualitative social science methods, design research methods and emerging digital methods. Governance in this context is defined as both top down and bottom up, and both within organisations and outside. We identify the role and impact of social expectations, high level guidelines, professional standards, (in)formal standards and norms, regulations and domain or organizational specific policies on the innovation dynamic of intelligent agents, investigated by way of policy research, public and expert engagement and user studies (where possible).

For example, the EU High-Level Expert Group on Artificial Intelligence has issued high level ethical guidelines for AI and put forward a human-centred approach on AI. One of the key goals is transparency but no specific approaches are given on how to achieve this. Transparency enables organisations, their clients, and those subject to AI decisions, to understand the decision making process and mitigate, or correct, potential harmful impacts. In order to enable meaningful transparency of future AI systems, a set of participatory and cooperative governance tools and resources need to be created that enable AI innovation but ensure that it is ethical and trustworthy. This requires the creation of new 'boundary objects'; i.e. material or organisational structures - functioning as information carriers, or bridges, that allow social groups from varying perspectives to develop a common language and work together (Heyman, 2018).

We have already created a range of boundary objects in previous projects. For example, in the DANDA[1] project, a *data collection form* aided AI engineers to identify data collection biases that are self-evident in the social sciences. In Synchronicity[2] a *GDPR-quiz* and *DPIA-threshold-questionnaire* were developed to translate legal knowledge. The quiz enabled smart city project partners to test their GDPR knowledge and learn to identify commonly made mistakes. The multiple choice questionnaire vided an indicator to a design team whether a Data Protection Impact Assessment was necessary or not. In SPECTRE[3], we are experimenting with design fictions (Lindley, 2016) that allow legal researchers to explore issues in future smart city projects. The difficulty in this case consists of creating legal documents, DPIA outcomes or technical descriptions of data processing operations that are specific enough to understand if a smart city project might be legal or illegal. Most project initiation documents are far too vague to perform a legal assessment and would not yield a conclusive legal answer. Finally, we have also explored the use of storytelling and theatre methods to develop cross disciplinary understanding of ethical and value challenges.

We are currently bringing our methods and diverse experiences together to develop a systematic approach to the development of cooperative responsibility in AI design and governance. In order to **operationalize** our approach, we have developed a three-phase strategy. First we investigate the divergence between high level guidelines and AI innovation in practice in specific domains to develop a roadmap of key social challenges that must be addressed. Second we identify and create intervention opportunities where these challenges can be addressed using boundary objects and co-creative practices that fit in the current project lifecycle of AI researchers and practitioners. Third we evaluate and assess the value and practical applicability of the developed tools and resources for AI practitioners and related stakeholders (including policy makers), and iteratively improve them based on feedback.

Based on the **outcome** of the three-phased strategy we plan to provide professional guidelines for those involved in developing and governing AI applications, and to develop a "public values dashboard". This dashboard will be used as a compass to indicate red lines and danger zones in AI-driven applications in the domain of media. The latter refers to instances where the application of future AI may lead to situations that transgress a threshold of one or more public values, or fundamental rights, that is not acceptable and/or allowed from a societal, ethical or legal perspective in Europe.

References

Ausloos, J., Heyman, R., Bertels, N., Pierson, J., & Valcke, P. (2018). Designing-by-Debate: A Blueprint for Responsible Data-Driven Research & Innovation. In F. Ferri, N. Dwyer, S. Raicevich, P. Grifoni, H. Altiok, H. T. Andersen, Y. Laouris, & C. Silvestri, *Responsible Research and Innovation Actions in Science Education, Gender and Ethics* (pp. 47–63). Springer International Publishing.

Helberger, N., Pierson, J. & Poell, T.(2018) 'Governing online platforms: from contested to cooperative responsibility', in *The Information Society*, 34 (1), 1-14.

Heyman, R. (2019). Sharing Is Caring, a Boundary Object Approach to Mapping and Discussing Personal Data Processing. In: Kosta, Eleni, Pierson, Jo, Slamanig, Daniel, Fischer-Hübner, Simone and Krenn, Stephan (eds.) *Privacy and Identity Management: Fairness, Accountability, and Transparency in the Age of Big Data*, Cham: Springer, 21-31.

Kerr, A., Barry, M., and Kelleher, J.D. (2020) In press "Expectations of Artificial Intelligence and the Performativity of Ethics: Implications for Communication Governance." *Big Data and Society*.

Lindley, J. (2016). A Pragmatics Framework for Design Fiction. *11th EAD Conference Proceedings: The Value of Design Research*. European Academy of Design Conference Proceedings, July 5, 2015.

Sloane, M. & Moss, E. (2019) 'AI's social sciences deficit', in *Nature - Machine Learning*, 1, August, 330-331.

Susser, D., Roessler, B. & Nissenbaum, H. (2019) 'Technology, autonomy, and manipulation', in *Internet Policy Review*, 8 (2).

[1] DANDA, Diverse and Accountable algorithm design, was an imec-internal project aiming to include accountability into machine learning development processes.

[2] https://synchronicity-iot.eu/.

[3] https://spectreproject.be/.