Towards an "Ethics by Design" methodology for AI research projects

Mathieu d'Aquin[1], Pinelopi Troullinou[2], Noel E. O'Connor[3],
Aindrias Cullen[3], Gráinne Faller[4], and Louise Holden[4]

[1] Insight Centre for Data Analytics, NUI Galway, Ireland, mathieu.daquin@insight-centre.org
[2] Knowledge Media Insititute, The Open University, UK, pinelopi.troullinou@open.ac.uk
[3] Insight Centre for Data Analytics, Dublin College University, Ireland, noel.oconnor@insight-centre.org,
aindrias.cullen@insight-centre.org
[4] FH Media Consulting Ltd, Dublin, Ireland, grainne@fhmediaconsulting.com, louise@fhmediaconsulting.com

Abstract

Addressing ethical issues arising from AI research, and by extension from most areas of Data Science, is a core challenge in both the academic and industry worlds. The nature of research and the specific set of technical skills involved imply that AI and Data Science researchers are not equipped to identify and anticipate such issues arising, or to establish solutions at the time a specific research project is being designed. In this paper, we discuss the need for a methodology for ethical research design that involves a broader set of skills from the start of the project. We specifically identify, from the relevant literature, a set of requirements that we argue to be needed for such a methodology. We then explore two case studies where such ethical considerations have been explored in conjunction with the development of specific research projects, in order to validate those assumptions and generalise them into a set of principles guiding an "Ethics by Design" method for conducting AI and Data Science research.

Introduction[1]

By its nature, research in AI and Data Science rely on large amounts of data, potentially impacting many different sectors and aspects of human lives. While research ethics is a longstanding area that has taken the role of ensuring that proper codes of conducts are in place for research activities that directly impact people in a potentially harmful way, and despite the fact that it is being increasingly discussed both in academia and in the general media, clear approaches and guidelines for data ethics in AI and Data Science research are still lacking. This is mostly due to the complexity of those research initiatives, where each project might have a wide, very varied set of implications, all associated with intricate societal and ethical risks. This challenge is further emphasised by the disconnect between the high technicality of research projects in AI and Data Science, and the need to introduce a social science perspective to ethics considerations in those projects.

The objective of this paper is therefore to offer an initial discussion, as a starting point toward establishing such approaches and guidelines, in the form of an "Ethics By Design" methodology for AI and Data Science research. We start by reviewing some related initiatives that aim to provide initial guidance supporting specific ethics-related processes (approval) or specific ethical issues (privacy). We derive from this initial requirements which we illustrate with two case studies on AI/Data Science projects where ethical considerations have been explicitly explored. We derive from both this review and the case studies a set of fundamental principles that should be implemented within a data ethics methodology guiding this kind of projects, and conclude on how those principles represent a starting point towards establishing a more concrete methodology.

Related Work

In this section, we review ongoing initiative that relate to the objective of creating an Ethics by Design approach to research in AI and Data Science. We start by reviewing the

work in the closely related area of Privacy by Design (although it is mostly concerned with software engineering), and show examples of ongoing initiatives generally related to the areas of research ethics and data ethics.

Privacy by Design: Privacy by Design is a framework, relying on a set of foundational principles used to embed privacy protection within the development lifecycle of (generally software) systems (Gürses et al, 2011). As the name suggests, these principles include the idea that privacy protection and the consideration for potential privacy issues are initiated and integrated from the design phase of the system (Cavoukian, 2010). It also includes the principles of privacy by default, that privacy issues should be anticipated in a proactive way, and not left to be dealt with in a reactive way, and more recently, that it relies on data minimisation (van Rest et al., 2012). Privacy by Design has however mostly been applied and understood as a framework for software/system development, and has therefore not been widely adopted within research methodologies. It also mostly focuses on the aspects of data protection, i.e. on avoiding unintended and unwanted access to private data, which from the point of view of data ethics, remain very limited. Interestingly too, while the two areas are rarely formally connected, the idea of embedding privacy considerations within the design and development lifecycle of a system naturally relates and relies on the notion of *Privacy Impact Assessment* (Wright and De Hert, 2012) which is the process of evaluating the ways in which specific projects could impact on the privacy of stakeholders. It often consists in evaluating and clearly communicating the privacy risks associated with the project in order to clarify the mechanisms by which those risks can be minimised or countered.

Research ethics guidelines: Strongly related to our interest in this paper is the area of research ethics, which has been a core academic concern both from a research and a practical point of view. This is especially prominent in areas of science with direct contact and consequences on humans (and to some extent animals) including the social sciences and medicine. Most of these areas adhere to long established codes and guidelines for ethical conduct in research (see for example the ones mentioned in Emanuel et al., 2008 for ethics in clinical research). Beyond general codes that are sometimes inherent parts of the inclusion within a specific research profession, processes need to be put in place to support researchers in assessing the potential ethical impact of their research, and ensure that the research is designed in such a way that its realisation will not contradict accepted ethical codes of conducts. We can cite for example (Guillemin and Gillam, 2004) which, in a spirit similar to the one of this paper, aim to introduce

reflexivity as a "*framework for thinking through [ethical] issues*". Furthermore, in communication studies, Potter box is also used as a practical way of reflection for practitioners to reason their decisions (see for example Backus and Ferraris, 2004; Watley, 2014). More pragmatically, most universities with significant research activities especially at least in the social sciences, medicine and biology areas have introduced mechanisms to validate and approve research projects from an ethics point of view.[2] These processes generally involve a phase of self-assessment by the researchers of the relevance of ethics issues, and in case there is relevance, the review by a dedicated "ethics committee" to decide on the ethical validity and interventions required in the research. Besides being by nature broad and not very detailed, we see two issues with those kinds of institutional processes when applied more generally to data ethics in AI and Data Science research: 1- They are only involved prior to the start of the research, resulting, in most cases, in binary results (i.e. the research can or cannot be conducted), and 2- They are only considering the ethical issues that might arise from conducting the research, and not the potential ethical consequences of the outcomes of the research. In many AI and Data Science projects, the research itself is perfectly ethical, but leads to developments and research products having potentially problematic, from an ethics point of view, societal effects. Indeed, despite all the work in ethics in qualitative studies, there have still been recently calls for more specific guidelines to support researchers (e.g. in Sanjari et al. 2014). Furthermore, the consent at the level of of both the participants and the users often acquired as response to privacy issues has become problematic. Advancements in Artificial Intelligence and the interconnection of different technologies and consequently different databases, can result in unexpected, even for the researchers, findings but also practices thus, there is an urge to redesign and reconceptualise the consent mechanisms (Luger and Rodden, 2013).

Data ethics - guidelines and recommendations: In contrast with the mostly academia-led area of research ethics, the emerging notion of data ethics has been driven over the last few years by a wide range of actors, without yet achieving a sufficient level of formalisation and codification. For the purpose of this paper, we define data ethics as *the set of principles and processes that guide the ethical collection, processing, analysis, use and*

---

[2] see for example
https://www.nuigalway.ie/research-office/policiesandprocedures/researchethicscommittee/,
http://www.open.ac.uk/research/ethics and
https://www4.dcu.ie/researchsupport/research_ethics/guidelines.shtml at the universities of the authors of this paper.

application of data having an effect on human lives and society. Unsurprisingly, much of the expectations so far regarding data ethics have been targeted to large technology companies, who have started responding by communicating their specific policies especially regarding their use and processing of Big Data.[3] This is also illustrated in government initiatives to, on the one end, explore the new ethical issues raised by big data (see e.g. the House of Commons Science and Technology Committee, 2014 report on the "Responsible Use of Data") and on the other end, setup high-level policies and guidelines for data ethics (see for example the UK's Cabinet Office, 2016 "Data Science Ethical Framework"). Those initiatives however tend to focus on privacy issues, and strongly relate to the upcoming EU General Data Protection Regulation. Following those emerging initiatives and debates happening outside of the academic circle, scholars in both technology-related and social science- or law-related areas are now exploring the various ethical challenges associated with the design, development and use of AI and Data Science projects (see for example Mittelstadt and Floridi, 2016; Richards and King, 2014). There is however consensus that such exploration cannot be achieved through the various disciplines involved (computer science, social sciences, law) working in isolation. For this reason, broader initiatives than the specific research endeavours mentioned above have started appearing in the last few years. Amongst those, in Ireland, we can mention the work carried out on data ethics by the ADAPT centre (see Reijers et al., 2016), as well as our own "Magna Carta for Data" project, which has been evolving from a the aim to create a general data ethics charter to a channel through which data ethics conversations can take place between the various stakeholders, in Ireland and beyond.[4]

Requirements for an "Ethics By Design" Methodology

From the sets of initiatives, methods and guidelines briefly reviewed above, we can extract common references to the need for a methodology for ethics in AI and Data Science research. Inspired by the Privacy-by-Design label, we aim for this work to expand from the often restricted notion of privacy to include a broader reflexivity regarding the research in AI and Data Science. As the name indicates, this implies bringing the debate on the ethical and societal implications at the primary stage of the research process

---

[3] see for example at IBM
https://www.ibm.com/blogs/policy/dataresponsibility-at-ibm/
[4] http://magnacartafordata.org

(see similar suggestions for a philosophically informed debate on ICT for transport Herzogenrath-Amelung, Troullinou and Thomopoulos, 2015)*, in a *proactive rather than a reactive way*. As clearly emerging from the initiatives reviewed above, a strong requirement for such a methodology is that it needs to include a *varied set of skills and expertises* in particular through enabling ethical challenges to be addressed as a conversation between computer scientists/technologists and social scientists/legal experts, in sharp contrast with the way in which AI and Data Science research is currently being conducted. Looking at the gap in current research ethics and privacy research, it is raised as a strong requirement for such a methodology to take a *broad view of data ethics*, not restricted to data protection or to the specific direct impact of the research at the time it is conducted.

In the next two sections, we illustrate those requirements and the need for such an Ethics by Design methodology to conduct research in AI and data science through two case studies. Both case studies are actual projects, one completed and another ongoing, where specific ethical issues have been raised through the involvement of social scientists. In both cases, the objective is to show what kind of issues might emerge from those research endeavor, and the gaps that still remain in the approach taken in addressing those issues in a way which is directly embedded within the research design.

Case Study in Data Ethics: Reverie

REVERIE was a 4 year project funded by the European Commission that aimed to develop technologies that would allow social networking to become immersive collaborative environments that support realistic inter-personal communication (Wall et al., 2014). To this end, the project integrated technologies enabling users to create virtual worlds and experience these world via virtual autonomous avatars (see Figure 1). It targeted technologies related to 3D data acquisition and processing, sound processing, autonomous avatars, networking, real-time rendering, and physical interaction and emotional engagement in virtual worlds.

The project was strongly focused on technology development. However, the potential ethical implications of introducing autonomy into virtual environments were recognized and reflected with a dedicated task within the project's work-plan to consider these aspects (Ó Brolcháin et al., 2016). The resulting study identified issues related to privacy and autonomy as the two most important areas to consider. The study discussed the various ways in which different concepts of privacy could be threatened e.g.

misuse of data (information privacy), loss of anonymity (physical privacy), and online socialising (associational privacy). In a similar manner ethical considerations regarding autonomy considered the concepts of knowledge, freedom, and authenticity, identifying threats such as filter bubbles, addiction and social conformity respectively. The study concluded by providing recommendations for policy-makers, providers and end-users.

Whilst recognized as extremely important in the context of the overall project, it could be argued that the treatment of ethical concerns followed a very traditional model and was thus, somewhat dis-connected to the thrust of technology development. That is, the study was carried out in parallel to the various work streams focused on definition of use cases, elicitation of functional and technical specifications and prototype development and validation. Also, the timing of the ethical work was such that the key conclusions and recommendations were only available towards the end of the project's lifetime. Thus, we argue that even though a debate on the ethical challenges had taken place, it was still situated outside the 'technical' scope of the research project. In contrast, we suggest that such a debate should be interconnected to the design of the system.

Figure 1: Virtual world created within the Reverie project.

Case Study in Data Ethics: Smart Stadium

The Smart Stadium project is a collaboration between Dublin City University (DCU), Intel, Microsoft and the Gaelic Athletic Association. Its aim is to develop an open accessible Internet-of-Things (IoT) test-bed to accelerate the development of innovative Smart City technology solutions (Panchanathan et al., 2017). The project uses Croke Park, the 3rd largest sports stadium in Europe as the venue for the test-bed, based on the hypothesis that a stadium represents a microcosm of a city. A stadium faces many of the challenges faced by cities, e.g. traffic, parking, human safety, energy efficiency, etc., at a scale sufficient to stress-test technology solutions whilst facilitating deployment of such solutions in practice. A stadium such as Croke Park meets the key criterion for a test-bed as a venue that is "small enough to trial, large enough to prove" technology solutions. Over the last two years the project has installed a state-of-the-art IoT test-bed whereby a large number of different types of sensors have been installed, connected to Intel IoT gateways, which in turn feed data to the Microsoft Azure cloud platform. A number of use-cases have been identified, including sound monitoring, crowd behavior understanding, turf care, and demonstrator systems have been built to investigate the benefits of IoT-enabled solutions.

Recognising the ethical and data privacy challenges that emerge from any IoT deployment, an ethical study was carried out by the Institute of Ethics in Dublin College University in collaboration with colleagues in Arizona State University (itself developing a similar smart stadium initiative) early in the project's lifetime. The study, the results of which are yet to be published in full, identified beneficial opportunities for the public but also ethical challenges. Opportunities included enhanced entertainment, improved customer service & commercial opportunities, enhanced safety & security, reduced environmental impacts & energy costs, and even improved human performance. The study classified the challenges into two categories. The risk of violation of ethical values included considerations such as privacy violations due to the ubiquitous nature of the sensing being performed and ambiguities relating to informed consent, among others. The risk of future negative consequences included potential issues such as possible misuse of sensitive information and challenges around allocation of responsibility.

The fact that the ethical study was carried out and completed early in the project's lifecycle was clearly of significant benefit to the researchers working on novel technology solutions. It helped facilitate ethically informed technology development for certain use-cases and demonstrators. For example, it was responsible in part for motivating investigations into edge-based processing of IoT data, ensuring that sensitive data could be processed securely on edge devices resulting in inherently anonymous data that could be forwarded on for further processing in the cloud. Conversely, the fact that the ethical work was completed at an early stage and not further progressed in parallel with the ongoing evolution of the project could be considered to be a missed opportunity. As is often the case, this was due to resource constraints. However, the project is extremely fortuitous that one of the

lead scientists has a background in anthropology ensuring deep consideration is given to ethical and privacy issues on an ongoing basis

Towards an Ethics by Design Methodology for Research Projects in AI

The two case studies of research projects presented above confirm to a large extent the need for a new methodological approach on ethical and societal concerns. They also depict the existing operational approach on ethics within AI and Data Science research. Based on these findings, the shortcoming of existing approaches identified above, and the experience provided by those two cases studies, we identify four general guiding principles towards an Ethics by design methodology.

Data ethics addressed as a conversation between technology and social aspects, all along the research process: There is no established consensus on what are the phases of the research process, and those very much vary from discipline to discipline. In Figure 2 however, we provide a minimal and reasonably uncontroversial set of steps for the purpose of this discussion. As can be seen from this figure, in most cases, institutional ethics approval processes tend to happen at the design stage, with the purpose of validating that the research activities established in the plan will not in themselves generate unresolveable ethical issues. Besides ethical approval being limited to a binary decision (approved/not approved), when considered against the experience provided through the case studies above, we can see that there are strong limitations especially in these cases where parts of the research plan includes technology-related development. Indeed, by restricting itself to the research design phase, it removes the ability for researchers to address ethical issues emerging from later phases, especially from specific results obtained, as well as from cycles of technology development, in any other way than by being reactive. This is especially visible in the "Smart Stadium" case study where, even though the data ethics-related conversation affected the research design as well as the development of the project itself in terms of technology, further results and later issues related for example with abuses of the collected information for unforeseen purposes will only be considered as they appear, and not proactively addressed.

Furthermore, the case studies show that a methodology exploring data ethics in AI and Data Science projects requires a varied set of skills, knowledge and perspectives. Indeed, in those cases, the involvement of social scientists with a purpose much broader than approving the project, namely to analyse the research plans so to explore its

possible implications, raised issues that might not have been considered by the technologists alone. In the case of Smart Stadium, this "conversation" led to changes in the technology-related development of the project.
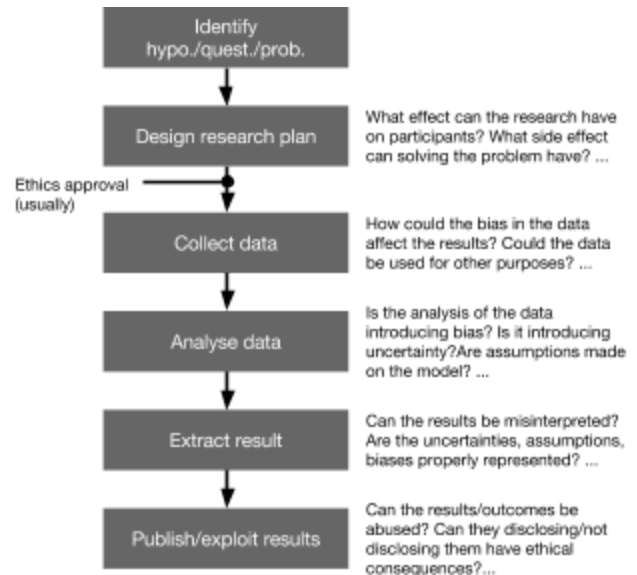

Figure 2: Stages of the research process, when ethical approval happens, and examples of questions to be asked at the different stages.

Ethics issues identified and addressed through reflexivity and anticipation: One of the issues with the research process presented in Figure 2 is that, without a proper handling of potential ethical issues, it is focused, designed and evaluated purely from the point of view of the hypothesis, research question or problem identified. This is the main reason why, in the previous paragraph and to a large extent in the case studies, the exploration of data ethics is presented as a conversation between social scientists and technologists. Indeed, ethics is not usually a criteria by which the performance of a research project is evaluated, which leads to ethical considerations being often seen by researchers as of lower priority than other aspects of the research. Those considerations should take place at a meta-process level, as they are related to the reflection researchers need to have upon their own research activities. Once again, such reflections (supported by research in the social sciences) need to be proactive rather than reactive. In other words, to say it plainly, it should be a constantly renewed process of asking "What will I do and how will it affect others/society?" rather than "What have I done and how can I fix it?"

The challenge here is that, with projects in AI and Data Science, the data collected and analysed, the results obtained and the way they can be used are often very complex and intricate, making it difficult to foresee consequences of specific activities. The particular form

taken by the reflection on data ethics in a research project in those domains should therefore be one of anticipation.[5] As illustrated in Figure 2, the questions to be asked at each stage all relate to what *could* happen or be done with the data and results of the research. In the case studies, the issues raised are not about actually materialised ethical issues, but potential privacy/data protection breaches and other aspects that could arise from the research (results).

Covering all aspects of the research: One of the issues with basic ethical approval is that it only covers the direct ethical issues that might emerge from conducting the research. Similarly, approaches such as privacy by design, only look into the initial phases of a project, and by extension tend to limit themselves to the direct consequences of what is being achieved. However, as it clearly appears in the case studies, first, many of the potential ethical consequences of AI and Data Science research do not emerge from the research itself, but from its results and even more from their exploitation (reinforcing the previous points related to covering the whole lifecycle of research). Indeed, taking the example of predictive analytics that often relies on AI approaches arround machine learning, even without considering how the data being collected might introduce biases, the inferences made by the prediction technique might lead to discriminations through underrepresenting and being more inaccurate for minorities groups. Generally, the principle to be followed here is that data ethics should consider not only the research activities, but also their impact; not only collected data, but also inferred information; not only privacy, but also autonomy, governance, fairness and equality; not only living human individuals, but also social groups, animals and the virtual personas of those passed away; not only our current society, but also the forming of future ones.

Impactful, not disruptive: A common objection to giving thorough consideration to ethics in research that does not directly impact on humans and animals is that it might slow down progress and prevent the research from happening. This is a valid argument when considering that ethics is achieved through an administrative process separated from the research itself. However, this argument should become invalid once the methodology for data ethics in AI and Data Science is embedded in the research itself. The methodology we suggest is a collaborative process part of the research project productive rather than disruptive. In the Reverie case study for example, the involvement of social scientists did not prevent the

research from happening, but enabled identifying potential risks that could be alleviated if and when the developed technologies was being deployed. Technology researchers in this project recognised the benefits of this collaboration, and that it could have had more value if started sooner. This is even more visible in the Smart Stadium case where, at least on some aspect, solutions could be found that reduced the risk of ethical issues appearing without diminishing the ability of the researchers to achieve their primary goal (and actually even adding value to it). In other words, the key principle here is that a methodology for data ethics in AI and Data Science is not about stopping the research when direct solutions are not found. It is, similarly to other approaches to risk management, about assessing the risks of ethical issues appearing and their potential impact, and establishing plans to minimise both those risks and impacts.

Conclusion

In this paper, we looked at approaches to address ethical issues in AI and Data Science research projects that are often associated with the manipulation of large amounts of real life data. We reviewed existing initiatives related to methods for addressing those issues and presented two case studies in which ethical considerations had been particularly emphasised, and extracted from both those cases studies and the extracted requirements four guiding principles for data ethics in AI and Data Science research:

1. cover the whole lifecycle of research,
2. be based on multidisciplinary reflexivity and anticipation,
3. cover all the aspects of the research, and
4. thrive to achieve positive impact, rather than disrupting research.

This discussion and the four principles elicited are viewed as a starting point towards a methodology for "Ethics by Design" in AI and Data Science research. Completing this methodology however will require further work on the specific approaches to implement the stated principles, including for example the use of "visual vignettes" to enable reflexivity and anticipation (see for example Troullinou et al., 2017). It will also require a broader collection of case studies from which to draw further principles and specific methods. Such a collection is currently being created based on crowdsourcing case studies through the MagnaCartaForData.org website.

---

[5] The word *anticipation* is used in french as a name for *social science fiction* which strongly relate to the kind of process being described here.

References

Backus, N. and Ferraris, C., 2004. Theory meets practice: Using the potter box to teach business communication ethics. *Association for Business Communication*, pp.222-229.

Cabinet Office (2016), Data Science Ethical Framework, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/524298/Data_science_ethics_framework_v1.0_for_publication__1_.pdf

Cavoukian, A. (2010). Privacy by design: The 7 foundational principles. implementation and mapping of fair information practices. Information and Privacy Commissioner of Ontario, Canada.

Emanuel, E. J., Grady, C. C., Crouch, R. A., Lie, R. K., Miller, F. G., & Wendler, D. D. (Eds.). (2008). The Oxford textbook of clinical research ethics. Oxford University Press.

Guillemin, M., & Gillam, L. (2004). Ethics, reflexivity, and "ethically important moments" in research. Qualitative inquiry, 10(2), 261-280.

Gürses, S., Troncoso, C., and Diaz, C. (2011). Engineering privacy by design. Computers, Privacy & Data Protection.

Herzogenrath-Amelung, H., Troullinou, P., and Thomopoulos, N. (2015). Reversing the order: towards a philosophically informed debate on ICT for transport. *ICT for Transport: Opportunities and Threats, 205.*

House of Commons Science and Technology Committee (2014), Responsible Use of Data, Fourth Report of Session 2014-2015

Luger, E., and Rodden, T. (2013, September). An informed view on consent for UbiComp. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing (pp. 529-538)*. ACM.

Mittelstadt, B. D., & Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. In The Ethics of Biomedical Big Data (pp. 445-480). Springer International Publishing.

Ó Brolcháin F, Jacquemard T, Monaghan, D.S, O'Connor N.E, Novitzky P. Gordijn B (2016), The convergence of virtual reality and social networks: threats to privacy and autonomy, Science and Engineering Ethics, 22 (2). pp. 1-29. ISSN 1471-5546

Panchanathan, S., Chakraborty, S, McDaniel, T., Tadayon, R., Fakhri, B., O'Connor, N. E., Marsden, M., Little, S., McGuinness, K., Monaghan, D. (2017) Enriching the fan experience in a smart stadium using internet of things technologies. International Journal of Semantic Computing 11(2)

Reijers, W., Gordijn, B. and O'Sullivan, D (2016), The Ethics of ADAPT, ADAPT Centre White Paper, https://www.adaptcentre.ie/images/uploads/about/Ethics_of_Adapt.pdf

van Rest, J., Boonstra, D., Everts, M., van Rijn, M., & van Paassen, R. (2012, October). Designing privacy-by-design. In Annual Privacy Forum (pp. 55-72). Springer Berlin Heidelberg.

Richards, N. M., & King, J. H. (2014). Big data ethics, Wake Forest Law Review

Sanjari, M., Bahramnezhad, F., Fomani, F. K., Shoghi, M., & Cheraghi, M. A. (2014). Ethical challenges of researchers in qualitative studies: the necessity to develop a specific guideline. Journal of medical ethics and history of medicine, 7.

Troullinou, P., Tiddi, I., and d'Aquin, M. (2017), Proceedings of the Re-coding Black Mirror 2017 Workshop co-located with 16th International Semantic Web Conference (ISWC 2017), CEUR-WS proceedings, http://ceur-ws.org/Vol-1939/

Wall J, Izquierdo E, Argyriou L, Monaghan D.S, O'Connor N.E, Poulakos, S, Smolic, A, Mekuria R, 2014, REVERIE: Natural human interaction in virtual immersive environments, IEEE International Conference on Image Processing (ICIP), 2165-2167, 2014

Watley, L.D., 2014. Training in ethical judgment with a modified Potter Box. *Business Ethics: A European Review*, 23(1), pp.1-14.

Wright, D., & De Hert, P. (2012). Introduction to privacy impact assessment. In Privacy Impact Assessment (pp. 3-32). Springer Netherlands.