Responsible and representative multimodal data acquisition and analysis: on auditability, benchmarking, confidence, data-reliance & explainability

Alice Baird, Simone Hantke, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Baird, Alice, Simone Hantke, and Björn Schuller. 2019. “Responsible and representative multimodal data acquisition and analysis: on auditability, benchmarking, confidence, data-reliance & explainability.” arXiv. Augsburg: Universität Augsburg. https://doi.org/10.48550/arXiv.1903.07171.

Responsible and Representative Multimodal Data Acquisition and Analysis: On Auditability, Benchmarking, Confidence, Data-Reliance & Explainability

Alice Baird[1], **Simone Hantke**[12], **Björn Schuller**[13]
[1]ZD.B.Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Machine Intelligence and Signal Processing Group, Technische Universität München, Germany
[3]GLAM – Group on Language, Audio and Music, Imperial College London, UK
alice.baird@informatik.uni-augsburg.de

Abstract

The ethical decisions behind the acquisition and analysis of audio, video or physiological human data, harnessed for (deep) machine learning algorithms, is an increasing concern for the Artificial Intelligence (AI) community. In this regard, herein we highlight the growing need for responsible, and representative data collection and analysis, through a discussion of modality diversification. Factors such as Auditability, Benchmarking, Confidence, Data-reliance, and Explainability (ABCDE), have been touched upon within the machine learning community, and here we lay out these ABCDE sub-categories in relation to the acquisition and analysis of multimodal data, to weave through the high priority ethical concerns currently under discussion for AI. To this end, we propose how these five subcategories can be included in early planning of such acquisition paradigms.

Keywords: Ethics, Multimodal, Representation, Data Acquisition, Data Analysis, Machine Learning

1. Introduction

Ethics itself is an encompassing term, and at its core it is the personal and societal decision making process between what is 'right' and what is 'wrong'. Although fundamental, within the research community the ethical (or moral) code of conduct can require quite some disentanglement, especially regarding Artificial Intelligence (AI) (Allen et al., 2006). Worldwide there are many non-technical groups of differing domains, which put ethics to the forefront of their manifesto, such as, the National Institutes of Health (NIH), or the National Science Foundation (NSF) (Resnik,). Additionally, some of the largest technology companies, such as Google's DeepMind are also bringing ethics to the forefront [1], and multi-organisation ethics boards are being developed, such as the Partnership on AI [2].

Herein, when referring to ethics, we are namely focusing on the field of AI and in particular machine learning. The *Ethics of AI* (Boddington, 2017) has become a 'hot topic' for researchers both internal and external to the AI community, with the peer-review process for publications itself being placed under scrutiny (Prechelt et al., 2018). The *Ethics of AI* and *Machine Ethics* (Baum et al., 2018) are quite different terms for the field of AI. The latter refers to giving conscious ethical based decision making power to machines. The former although somewhat informing the latter, refers more broadly to decisions made by the researchers involved, covering diversity and representation, e. g. to avoid discrimination (Zliobaite, 2015), or inherent latent biases that may come from archival databases (van Otterlo, 2018). As AI is largely data-driven, understanding its environment through context provided via annotations, *Big Data* ethics for AI algorithms is an expanding discussion point (Berendt et al., ; Mittelstadt and Floridi, 2016). In this regard, crowdsourced data (i. e. data gathered from large amounts of paid or unpaid individuals via the internet) – made by humans with little time and minimal interest in the annotation task – is one aspect which is raising concern, due to the resulting validity of such annotations (Hantke et al., 2016).

Annotators are only limited by the classes provided to them by the researcher, and multi-modality during data labelling is only a technical challenge. A single domain model alone e. g. a speech recognition system, can limit the potential 'intelligence' of resulting AI algorithms, restricting the overall interaction. For truly 'social' AI interactions, multimodal aware data should be of high priority, along with the pressing concerns of safety and privacy (Fox, 2011). Multimodal data, in turn brings more complex interpretation concerns e. g. the term 'grasp' is both a natural language processing, and gestural recognition problem (Mangin, 2017). Multiple modalities also multiplies the need for data auditability, and efficacy (Tørresen, 2018).

2. Related Work and Motivations

With AI efforts increasing, the ethical demands related to the needed Big Data, are expanding in parallel (Herschel and Miori, 2017). A general consideration which is not being completely overlooked, particularly in the field of Natural Language Processing (Leidner and Plachouras, 2017). With a vast amount of data sourced online, e. g. through social media platforms, the legal aspects in terms of user privacy are at the forefront (Lanfranchi, 2017).

Machine Learning Fairness (i. e. bias), is currently a popular topic[3]. With three core biases discussed e. g. Interaction Bias, Latent Bias, and Selection Bias, in this paper we focus primarily on Selection Bias, as we propose that true multimodal and representative data could assist in avoiding this. Selection Bias i. e. the by-products of decisions during collection and analysis – including misrepresentation e. g. through unbalanced gender classification (Gao and

[1]https://deepmind.com/applied/deepmind-ethics-society/
[2]https://www.partnershiponai.org/board-of-directors/
[3]https://developers.google.com/machine-learning/fairness-overview/

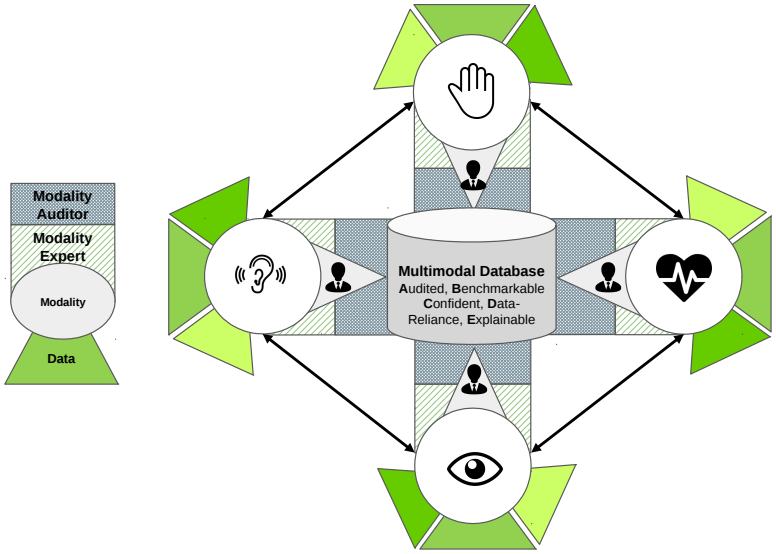

Figure 1: Proposed ABCDE approach for large-scale data collection across modalities. Considering multiple data sources, covering a spectrum of identities. Allowing for improved representation, as well as human safeguards within each modality.

Ai, 2009) – is an important concern for technology companies. Such biasses can quite easily propagate into a resulting system, making it not only ethically problematic, but also fundamentally producing commercial limitations, e. g. who does the model represent, and who will buy it?
Identity-representation itself is a prominent topic in AI today (Fussel, 2017), and like many other human traits its manifestation within AI can be limited. The warnings are quite realistic, as bias created from the developers themselves (Zliobaite, 2015), or through pre-existing archival (i. e. data from historic sources) data (van Otterlo, 2018), data-driven machine learning algorithms will simply replicate this. In this regard, the need for multimodal, representative data is more prevalent than ever. Not only, due to the aforementioned demographic biasses, including for gender-based variables (Larson, 2017), which can occur during collection, but also for representation and usability on a global scale, improving the overall impact of *intelligent* HCIs. As well as this, for effective Human Computer Interactions (HCIs) social robotics require observation of a scene through multiple modalities by for example enhancing the ability for gesture recognition (Pitsikalis et al., 2015). In the age of deep learning, multimodal interactions are the next stage for enhancing the usability of such algorithms (Gordienko et al., 2018; Baltrusaitis et al., 2017). Offering benefits across domains, including conditions with a broad variety in population needs, such as Autism (Liu et al., 2017b).

3. Considerations

With this motivation in mind, we now take a closer look at five core sub-categories which have been highlighted individually within the machine learning community, as factors which should be considered during data acquisitions and analysis, particularly as it pertains to multimodal data, as a means of informing a better identification representation within human data. Clearly, this represents only a very small sub-genre of the ethics within AI, yet of fundamental importance to the data-driven algorithm we quite commonly see, e. g. embedded mobile assistants, or self-driving cars. These sub-categories include, Auditability, Benchmarking, Confidence, Data Reliance, and Explainability, and Figure 1 visualises a potential protocol, which can ensure that during collection and resulting analysis these factors can be considered in a more systematic way.

3.1. Auditability

As the collection, annotation and analysis of truly Big Data can be extremely time consuming and costly for those involved, there are many methods now being developed to make collecting and annotating data in a semi-automatic fashion possible, including through social-media crawling (Amiriparian et al., 2017), and Active Learning annotation (Hantke et al., 2017). One consideration that engineers should keep in mind when designing such algorithms, is the balance between manual and automatic annotation. Having a human (manual) auditor present within such a database, is needed for many aspects from quality, to legality. Although the realisation may have technical challenges in the balance between public auditing and privacy (Diakopoulos and Friedler, 2017), the need for this has been highlighted in the literature in terms of security (Wang et al., 2011), and privacy (Tong et al., 2014), with some showing concerns that autonomous agents responsible for collecting data, e. g. directly from human manually labelled tags on YouTube, may breach privacy restriction or incorrectly store the resulting data due to an initial human error. In this regard, crowd-sourcing annotations should also be continually audited by human intervention with aspects such as fraud, being an ever present issue (Rothwell et al., 2015).

3.2. Benchmarking

Benchmarking between system elements has shown to be difficult for multimodal, single domain databases (Liu et al., 2017a). A clear and explainable comparison between resulting systems has many advantages for development and improvement, yet aspects such as specific modalities are continually miss-matched across databases i. e. in the image domain some may have RGB values, but miss the Skeleton extraction. In order to combine data, and fully utilise its potential, in the literature researchers have begun to closely look at benchmarking between multimodal

databases (Liu et al., 2011), in this way improving HCI interactions. Through a more considered and ethical protocol during multimodal data collection, researchers could be guaranteed the possibility of benchmarking against the data of others.

3.3. Confidence

Having 'confidence' in one's database, comes across through these deeper considerations, and in turn offers in some way a more accurate long term system, with enhanced moral understanding (Blass, 2018). However, here we refer more to the use of confidence as a measure i. e. how accurate is the current system prediction, as a means of understanding the current risk (Duncan,), additionally to aid future fault detection. This becomes specifically important when discussing AI systems which are designed for Human-Care, as not providing an overall confidence in data behind that, can result in a substantial risk to the human user (Ikuta et al., 2003). Estimation of such attributes of data has been a focus for many researchers, with some also suggesting that the implementation of a 'self-confidence' for social robotics could be a strong aid in avoiding accidents during interaction (Roehr and Shi, 2010). Through multimodal near human-like perceptive data, confidence could be substantially improved across a system, as missing a single modality e. g. audio domain, is an over all a weakness for the system, creating a blind-spot.

3.4. Data-Reliance

Similar to confidence, data reliability is the process of completing acquisition of a database without error, within the context of the domain it is targeted towards (Morgan and Waring, 2004). To show such reliability, there are standardised statistical tests such as p-values, which can be used as evidence of significance for particular aspects of the database. However the use of such tests, has begun to gain criticism in recent years, due to their extensive misuse by the machine learning community (Vidgen and Yasseri, 2016). Reliability of data is also discussed as a basis for the Internet of Things (IoT), as often (cf. Section 3.1) data is being sourced through social media (amongst other online sources) for the IoT, and researchers have suggested that such a single modal application may weaken the user experience.

3.5. Explainabiltity

In recent years the machine learning community has been dominated by the need for accuracy, and a competitive nature has spread throughout the field, with internationally open challenges (asking participants to improve on a baseline system of a particular database), such as the Interspeech Computational Paralinguistics Challenges (Schuller et al., 2013) and the ACM Audio / Visual Emotion Challenges (Schuller et al., 2012), assisting in the rapid advancements of approaches across multiple domains. Although healthy competitions and high accuracy has its advantages, one limitation to this is the lack of explainabiltiy across the field (Huszár, 2015). It is of up most importance that such machine learning models are interpretable, offering a clear use-case, as ultimately, without this, the results alone are potentially meaningless (Vellido et al., 2012). As machine learning is predominately a pattern recognition task, visualisation of data has been a key enhancement for system explainabiltiy, particularly in deep learning, as this allows for trends within data to be more easily understood and interpreted (Samek et al., 2017).

4. Concluding Remarks

The current by-products of machine learning algorithms in relation to selection bias have been discussed throughout this paper, placing five key ABCDE sub-categories as they pertain to the enhancement of acquisition and analysis for multimodal data-driven models. We summarise these five key sub-categories which should be a considered focus to any researchers in the field of AI and machine learning. With ethics groups being founded by many of the largest AI based companies, there is clearly momentum towards ethical considerations within the community. One giant step forward in this regard is the multi-organisation discussion such as the Partnership on AI, but to the best of the authors' knowledge, there is not yet a large scale interdisciplinary discussion of researchers across domains. This would be a necessary step forward by the engineering community to include researchers from differing backgrounds, as AI is seemly going to be embedded in every aspect of our daily life, of which culture, and the arts, enrich substantially. In this regard, multi-modality should also not be restricted to the fundamental layers of the human senses, but expanded to more subtle differences, making way for a richer human imitation. This may be more appropriately achieved through interdisciplinary discussion, manifesting a more human-like tapestry of possibilities. Additionally in machine learning, the push toward *0-shot*, or *unsupervised learning* techniques, could be something to be wary of, and perhaps an audited, (at least semi-)supervised learning approach, is best for long-term implementations. With this in mind, considering a cross-modal data source, the implications of the five sub-categories discussed herein can be a step forward to a more representative AI model.

5. Acknowledgements

This work is funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B), and the European Union's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu).

6. Bibliographical References

Allen, C., Wallach, W., and Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, 21:12–17.

Amiriparian, S., Pugachevskiy, S., Cummins, N., Hantke, S., Pohjalainen, J., Keren, G., and Schuller, B. (2017). CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms. In *Proc. of Int. Conference on Affective Computing and Intelligent Interaction*, pages 340–345, San Antonio, TX, USA.

Baltrusaitis, T., Ahuja, C., and Morency, L. (2017). Multimodal machine learning: a survey and taxonomy. *arXiv*, abs/1705.09406.

Baum, K., Hermanns, H., and Speith, T. (2018). From machine ethics to machine explainability and back. In *Proc. of Int. Symposium on Artificial Intelligence and Mathematics*, Florida, USA. no pagination.

Berendt, B., Büchler, M., and Rockwell, G.). Is it research or is it spying? thinking-through ethics in big data AI and other knowledge sciences. *Kunstliche Intelligenz*, 29(2).

Blass, J. A. (2018). You, me, or us: balancing individuals' and societies' moral needs and desires in autonomous systems. *AI Matters*, 3(4):44–51.

Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Springer International Publishing, Cham, Switzerland.

Diakopoulos, N. and Friedler, S. (2017). How to hold algorithms accountable. [online] http://bit.ly/2f8Iple.

Duncan, B.). Importance of confidence intervals. *Insights Association*. [online] http://bit.ly/2pgT4kM.

Fox, J. (2011). Safe, sound and ethical: Rolling out the intelligent systems of the future. *ERCIM News*, 2011:40–41.

Fussel, S. (2017). AI Professor details real-world Dangers of Algorithm Bias . [online] http://bit.ly/2GDoudz.

Gao, W. and Ai, H. (2009). Face gender classification on consumer images in a multiethnic environment. In *Proc. of Int. Conference on Advances in Biometrics*, pages 169–178, Alghero, Italy.

Gordienko, Y., Stirenko, S., Kochura, Y., Alienin, O., Novotarskiy, M., and Gordienko, N. (2018). Deep learning for fatigue estimation on the basis of multimodal human-machine interactions. *arXiv*, abs/1801.06048.

Hantke, S., Batliner, A., and Schuller, B. (2016). Ethics for Crowdsourced Corpus Collection, Data Annotation and its Application in the Web-based Game iHEARu-PLAY. In *Proc. of Int. Workshop on ETHics In Corpus Collection, Annotation and Application, satellite of the Language Resources and Evaluation Conference*, pages 54–59, Portoroz, Slovenia.

Hantke, S., Zhang, Z., and Schuller, B. (2017). Towards Intelligent Crowdsourcing for Audio Data Annotation: Integrating Active Learning in the Real World. In *Proc. of INTERSPEECH*, pages 3951–3955, Stockholm, Sweden.

Herschel, R. and Miori, V. M. (2017). Ethics & big data. *Technology in Society*, 49:31–36.

Huszár, F. (2015). Accuracy vs explainability of machine learning models. *inFERENCe*. [online] http://bit.ly/2GAfW7c.

Ikuta, K., Ishii, H., and Nokata, M. (2003). Safety evaluation method of design and control for human-care robots. *The International Journal of Robotics Research*, 22:281–297.

Lanfranchi, V. (2017). Machine Learning and Social Media in Crisis Management: Agility vs. Ethics. In *Proc of. Int. Conference on Information Systems for Crisis Response And Management*, Albi, France.

Larson, B. (2017). Gender as a variable in natural-language processing: Ethical considerations. In *Proc. of the ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain.

Leidner, J. L. and Plachouras, V. (2017). Ethical by design: Ethics best practices for natural language processing. In *Proc. of ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain.

Liu, H., Feris, R. S., and Sun, M. (2011). Benchmarking datasets for human activity recognition. In *Visual Analysis of Humans - Looking at People.*, pages 411–427.

Liu, A., Xu, N., Nie, W., Su, Y., Wong, Y., and Kankanhalli, M. S. (2017a). Benchmarking a multimodal and multiview and interactive dataset for human action recognition. *IEEE Trans. Cybernetics*, 47:1781–1794.

Liu, W., Zhou, T., Zhang, C., Zou, X., and Li, M. (2017b). Response to name: A dataset and a multimodal machine learning framework towards autism study. In *Proc. of Int. Conference on Affective Computing and Intelligent Interaction*, pages 178–183, San Antonio, TX, USA.

Mangin, O. (2017). Multimodal concepts for social robots. *AI Matters*, 3:19–20.

Mittelstadt, B. D. and Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics*, 22:303–341.

Morgan, S. and Waring, C. (2004). Guidance on testing data reliability. *City of Austin*. [online] http://bit.ly/2kjNgX4.

Pitsikalis, V., Katsamanis, A., Theodorakis, S., and Maragos, P. (2015). Multimodal gesture recognition via multiple hypotheses rescoring. *Journal of Machine Learning Research*, 16:255–284.

Prechelt, L., Graziotin, D., and Fernández, D. M. (2018). A community's perspective on the status and future of peer review in software engineering. *Information & Software Technology*, 95:75–85.

Resnik, D. B.). What is ethics in research & why is it important? *National Institue of Enviromental Health Sciences*.

Roehr, T. M. and Shi, Y. (2010). Using a self-confidence measure for a system-initiated switch between autonomy modes. In *Proc. of Int. Symposium on Artificial Intelligence, Robotics and Automation in Space*, pages 507–514, Sapporo, Japan.

Rothwell, S., Elshenawy, A., Carter, S., Braga, D., Romani, F., Kennewick, M., and Kennewick, B. (2015). Controlling quality and handling fraud in large scale crowdsourcing speech data collections. In *Proc. of INTERSPEECH*, pages 2784–2788, Dresden, Germany.

Samek, W., Wiegand, T., and Müller, K. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv*, abs/1708.08296.

Schuller, B. W., Valstar, M. F., Eyben, F., Cowie, R., and Pantic, M. (2012). AVEC 2012: the continuous audio/visual emotion challenge. In *Proc. of Int. Conference on Multimodal Interaction*, pages 449–456, Santa Monica, CA, USA.

Schuller, B. W., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K. R., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proc. of INTERSPEECH*, pages 148–152, Lyon, France.

Tong, Y., Sun, J., Chow, S. S. M., and Li, P. (2014). Cloud-assisted mobile-access of health data with privacy and auditability. *IEEE J. Biomedical and Health Informatics*, 18:419–429.

Tørresen, J. (2018). A review of future and ethical perspectives of robotics and AI. *Front. Robotics and AI*, 2018:75.

van Otterlo, M. (2018). Gatekeeping algorithms with human ethical bias: The ethics of algorithms in archives, libraries and society. *arXiv*, abs/1801.01705.

Vellido, A., Martín-Guerrero, J. D., and Lisboa, P. J. G. (2012). Making machine learning models interpretable. In *Proc. of European Symposium on Artificial Neural Networks*, Bruges, Belgium.

Vidgen, B. and Yasseri, T. (2016). P-values: misunderstood and misused. *arXiv*, abs/1601.06805.

Wang, Q., Wang, C., Ren, K., Lou, W., and Li, J. (2011). Enabling public auditability and data dynamics for storage security in cloud computing. *IEEE Trans. Parallel Distrib. Syst.*, 22:847–859.

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. *arXiv*, abs/1511.00148.