NC STATE UNIVERSITY
College of Humanities and Social Sciences
S. Douglas, N. Edgren, & V. Dubljević
Contact: veljko_dubljevic@ncsu.edu

Ethical Perspectives from AI Experts on Reverse-Engineering the Brain

Introduction

- The future of artificial intelligence is far from certain, but AI presents many immediate concerns and research opportunities for ethical decision making in AI (Bauer & Dubljević 2019)
- Using AI to reverse-engineer the brain may aid researchers in understanding both human and artificial intelligence, but there is little empirical ethics research into which areas are most relevant or feasible.
- As in other areas of neuroethics scholarship, there is a need to separate urgent from speculative issues (Racine et al. 2017)

Methods

- Building on prior work (Dubljević et al. 2021), we conducted nine semi-structured interviews with computer science and engineering experts specializing in AI near Raleigh, North Carolina
- We seek to discern the issues, concerns, and directions for future research in AI around nine broad themes, four from the R.L. Rabb Symposium (AI in Society 2021) and five from the National Academy of Engineering's Grand Challenges (NAE). Two of these themes were of particular interest:
  - **Integrating ethics in to AI decision making**
  - **Reverse-engineering the brain**
- We used two methods of data analysis in order to get a comprehensive grasp on what experts in the field felt were pertinent areas as AI advances:

Qualitative Interview Methodology (*Timmermans & Tavori 2012*)	
Data analysis conducted concurrently with data collections	
Codes were developed via abductive analysis, and inter-coder reliability was high (84.38%)	
Data vs. Algorithms	1. How decisions are made 2. Ethics of data sets 3. Echo chambers and misinformation
Reverse-engineering the brain	1. Viability of neural networks 2. In order to understand the human brain 3. In order to influence or manipulate human behaviors

Delphi Methodology (*Okoli & Pawlowski 2004*)	
First round (interviews)	• Asked interviewees to list at least three topics for each theme • Demographic data was collected (gender, disciplinary background, and location)
Second round (survey)	• Topics from the first round were consolidated into manageable, short sentences • Asked respondents to select three important topics and three that are not feasible
Third round (survey)	• Topics that were selected by a majority of respondents were included • Asked respondents to rank each topic based on importance, desirability, and feasibility

Qualitative Interview Quotes

Data vs. Algorithms	Reverse-engineering the brain
"Anybody who falls outside of those norms is going to have a higher chance of being misclassified. And falling outside of those norms is, of course, relative to the data set that it's trained on" (Page 6)	"One of the concerns we do need is, that everything becomes automatic, and on the one hand, it's good that it's 'humans hand-free,' but on the other hand, we left the human out of the book" (Page 3)
"Specifically, […] issues related to […] equity and fairness of the machine learning algorithms and also the data sets that actually training these algorithms, so you don't have […] a data set that is properly representative [...] of your population, […] then certainly this is going to cause some sort of bias on the algorithms" (Page 52)	"[T]here's a lot of excitement about artificial neural networks, but how closely these actually teach us anything about actual neurons is […] a real question that hasn't been addressed very well" (Page 33)

Delphi (First Round Topics)

Integrating ethics into AI decision making	Reverse-engineering the brain
Biases built into algorithms and training data	Leaving the human out of the picture
Which ethical theories to introduce or use in AI	Use AI as a neuroscience research aid (e.g. using AI for behavioral research or with EEGs to monitor sleep cycles)
Equity and fairness in machine learning and data sets	Use neuroscience research to aid AI research
Representational issues in the data inputted into AI	Brain-computer interfaces
False positives (identified by the AI) that lead to negative outcomes	Neurodiversity among human brains and/or embedding a hegemony of averages into a reverse-engineered brain

Conclusion

- Given the open-ended potential for AI to aid both human and artificial intelligence, it is imperative to understand the best practices in AI research and the implications for neuroethics.
- Our study shows that there is a strong desire among AI researchers to keep ethics and public good in mind when training AI systems to reverse-engineer the brain.

References

Bauer, W.A. & Dubljević, V. (2019). AI Assistants and the Paradox of Internal Automaticity, Neuroethics, 13:303-310. https://doi.org/10.1007/s12152-019-09423-6.
Racine, E., Dubljević, V., Jox, R., Baertschi, B., Christensen, J.F., Farisco, M., Jotterand, F., Kahane, G. & Müller, S. (2017):Can neuroscience contribute to practical ethics? A critical review and discussion of the methodological and translational challenges of the Neuroscience of Ethics, *Bioethics* 31(5): 328-337.National Academy of Engineering. (n.d.). *Reverse-Engineer the Brain*. NAE Grand Challenges for Engineering. http://www.engineeringchallenges.org/challenges/9109.aspx
Dubljević, V., Douglas, S., Milojevich, J., Ajmeri, N., Bauer, W.A., List, G.F. & Singh, M.P. (2021). Moral and Social Ramifications of Autonomous Vehicles, *ArXiv* http://arxiv.org/abs/2101.11775
AI in Society Group at NC State. (2021). *Symposium*. The AI in Society Group at NC State. https://sites.google.com/view/ai-society-at-nc-state/symposium?authuser=0
Timmermans, S., & Tavory, I. (2012). Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis. Sociological Theory, 30(3), 167–186. https://www.jstor.org/stable/41725511
Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, *42*(1), 15–29. https://doi.org/https://doi.org/10.1016/j.im.2003.11.002