Memento hominibus: on the fundamental role of end users in real-world interactions with neuromorphic systems

Serge Thill
Donders Institute for Brain, Cognition, and Behaviour
Radboud University
6525 HR, Nijmegen, the Netherlands
and
Interaction Lab, School of Informatics
University of Skövde
541 28 Skövde, Sweden

Maria Riveiro
Department of Computer Science
and Informatics, School of Engineering
Jönköping University
551 11 Jönköping, Sweden

Abstract

In this contribution, we briefly examine the role of end users in the evaluation and characterisation of sophisticated AI-based systems, such as autonomous vehicles or near-future robots. Indeed, when trying to ensure the safety of learning, perception and control in real world settings, one aspect that needs consideration is that human end users are often part of such settings.

We argue that current approaches for considering end users in this respect are insufficient, not the least from a safety perspective, and that this insufficiency will become more acute when transitioning to neuromorphic and/or strongly cognitively inspired solutions. We demonstrate this by borrowing examples from the field of enactivism, which demonstrate that human end users might change the system dynamics of advanced neuromorphic systems when interacting with them, which needs to be taken into consideration. Enactivism might also provide clues as to how to design future evaluation metrics for human-machine teams.

1 Introduction

Sophisticated machines are increasingly becoming a feature of modern-day society, and human collaboration with such machines increases correspondingly. Such machines will employ state-of-the-art machine learning algorithms, many of which are, at least in some form, neuromorphic. They can be so, at a minimum, by virtue of using deep neural networks or similar neural architectures, but evidence that more biologically plausible spiking networks are a viable solution, potentially running on dedicated neuromorphic hardware, is accumulating rapidly. For example, DeWolf et al. (2016) recently demonstrated adaptive control of a robot arm using spiking neural networks while Blouw et al. (2018) have shown that spiking neural networks running on Intel's Loihi platform can, for the right application, lead to significantly increased power efficiency without penalty in accuracy.

Neuromorphic approaches, some going beyond just deep learning, are also beginning to find applications in autonomous vehicles, whether it is for sensorimotor control using subsumption architectures (Plebe et al., 2019) or more generally a biologically inspired cognitive controller, including inspiration from human action selection and the ability to imagine hypothetical events (Da Lio et al., 2017).

This is therefore an opportune moment to reflect what this push towards increasingly biologically plausible and neuromorphic control, instantiated in machines that increasingly aimed at interacting with human end users in tasks that can be relatively complex, implies for how such systems can be evaluated, whether this is for safety, for ease of use, or other aspects.

The core argument in this contribution is that present-day approaches are not sufficient in this, because the neuromorphic aspect changes how information is integrated in the system. Specifically, we will argue, that the role of end users – used here for the lack of a better word – goes fundamentally beyond that of merely

Appearing in Proceedings of the Workshop on Robust Artificial Intelligence for Neurorobotics (RAI-NR) 2019, University of Edinburgh, Edinburgh, United Kingdom.

a user: they are, in effect, a non-abstractable part of the system itself. We will show this briefly by looking to cognitive science, specifically enactivism, as a field which has long studied interactions between neurally controlled agents (including, of course, humans, but also simple robots that are often used to demonstrate some cognitive mechanism). We conclude by arguing that it is also in these fields, rather than merely the technological ones, that we can find clues for how to approach the evaluation of physical neuromorphic systems built for interaction with humans.

2 Current approaches to evaluating human-machine collaborations

How to evaluate a human ML-system collaboration remains an unresolved challenge. Typical evaluations of humans interacting with ML-systems are done using traditional methods and metrics from the concerned communities, namely the machine learning (ML; algorithm-centered evaluation) and human-computer interaction (HCI; human-centered evaluation) crowds (see recent surveys of methods and metrics in Mohseni et al. (2018) and Hoffman et al. (2018)). This has led to a paucity of holistic and integrative methods that assess the overall collaboration over particular components (Boukhelifa et al., 2018).

ML normally uses performance evaluation metrics that focus on the algorithms presented, such as accuracy, precision, recall, squared error, likelihood, posterior probability, information gain, and so on; metrics that are also often used in robotics to evaluate the functioning of a system. Interactive ML (iML), meanwhile, improves on these metrics by typically combining them with some form of usability assessments (see, for example Talbot et al., 2009; Ribeiro et al., 2016; Cremonesi et al., 2011), *i.e.*, the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use. Such usability evaluations can be classified in exploratory, formative and summative evaluations.

Exploratory evaluations examine the current usage of a system, and normally use observations, interviews, surveys and automated logging. Formative evaluations help improve the system during the design process, employing heuristics and thinking aloud methods. Summative evaluation assesses the overall quality of a system once is more or less finished by collecting bottom-line data and quantitative measurements of performance: how long did users take, were they successful, how many errors did they make, number of commands/features used, etc.

HCI, meanwhile, interacts with other fields that have made noteworthy progress in the challenging evaluation of their interactive systems. One of these is the Visual Analytics (VA) community. Here, metrics include subjective user ratings, decision time, satisfaction, confidence, number of insights, accuracy, time on task, subjective preference, subjective choice assessment, ease and attachment (Dimara et al., 2018). part of the problem of designing and developing adequate evaluation methodologies resides in the fact that it is difficult to define concepts such as "insight" or "knowledge discovery"; even if definitions exist, there is not one that is commonly accepted by the research community (Yi et al., 2008).

It is also worth pointing out that some work is explicitly concerned with assessing the effectiveness of explanations provided by iML-systems. The use of explanations associated to ML-system outcomes has generally brought positive results (the use of explanations has been extensively studied in the context of knowledge-based systems, see Gregor and Benbasat, 1999), but it is not clear what information inherent to iML-systems such explanations should contain. Lim et al. (2009), for example, showed that explanations describing why the system behaved a certain way resulted in better understanding and stronger feelings of trust; but explanations describing why the system did not behave in a certain way resulted in lower understanding yet adequate performance. Other concerns or limitations of the use of explanations for calibrating trust or providing decision support have been raised recently by, e.g., Springer and Whittaker (2018); Dietvorst et al. (2015).

The evaluation of explanations and their effectiveness tend to be a combination of different HCI and ML methods as well, and as highlighted by the DARPA report on XAI: these are for example, user satisfaction through user ratings (clarity and utility of the explanation), trust assessment, correctability (identifying errors, correcting errors and continuous training), task performance (does the explanation improve the user's decision or task performance) and mental model (understanding individual decisions, understanding the overall model). There is thus a strong push towards a human-centred AI in the field (Biran and Cotton, 2017; Kirsch, 2018). However, even in these endeavours, humans are seen only as the end users of the system, and this is reflected in the metrics that are used for evaluation.

3 Roles beyond end users

The behaviour of a system trivially depends on humans in the sense that they are users of the system

and therefore provide inputs, if nothing else, to ensure that a task is achieved. Similarly, the way that a system interacts with a human influences their behaviour. In some of our own work, for example, we have seen that humans will rate driver assistance systems that give recommendations for certain actions (for example, which direction to take at a crossing or for eco-friendly driving behaviours) higher if these systems justify *why* these recommendations are given (Thill et al., 2014, 2018). This remains true even if the information from the system is factually wrong (for example, claiming that there is less traffic on a selected route even though the opposite is true Thill et al., 2014). Critically, these recommendations affect the behaviour of people, including possibly those not directly targeted by the recommendation. For example, we found that the perceived intelligence of a navigation aid modulated the time they spend looking through the front windscreen (Thill et al., 2014),

More generally, from such a perspective, one can consider the human-machine system to be a team (not restricted to just one human or one machine), and then evaluate the team based on its performance as a whole. This goes beyond the types of evaluations considered before, and it too is not trivial – one can for example question if machines should be considered full team members at all (Groom and Nass, 2007). Even if that question is ignored, different types of human-machine teams place different responsibilities on the different team members, which affects how one might analyse the overall behaviour (see Lagerstedt et al., 2017, for a discussion).

How to evaluate team of humans and robots has received attention in the past in human robot interaction (see, for example Dautenhahn, 2007; Breazeal, 2004), and it is arguably viable to look to the cognitive sciences for guidance (Thill and Ziemke, 2015). In enactive approaches to cognitive science in particular, agents are understood as coupled to the environment, and thereby also to other agents in this environment (Maturana and Varela, 1987). This then leads, amongst many others, to the question whether or not these interactions are actually crucial in order to understand the behaviour of a cognitive system, or whether this system can be analysed and understood, as it is in more computationalist cognitive science, by focussing solely on this system. For example, social cognition, although it is fundamentally about interaction with other agents, is – within these computationalist branches – primarily studied in terms of mechanisms at play in the heads of the individuals: the interaction that follows is merely an output of these processes. This is challenged by more enactive views, in which the interaction itself is hypothesised to be constitutive of social cognition, and not just merely an output (De Jaegher et al., 2010).

One of the key features of enactivism is that the mathematical framework of choice is dynamical systems theory. This has led to relatively simple demonstrations of the fundamental ideas in robotics – often, using simple wheeled robots that have basic neural network controllers. In this context, Froese et al. (2013) provide a noteworthy investigation into the role of the coupling wiht the environment. In their study, the authors simulate simple robots that are controlled by just one neuron. From mathematical principles, it is known that such a dynamical system cannot exhibit particularly interesting behaviour; it requires at least two dimensions for oscillatory behaviour, and three or more for chaotic behaviours.

Nonetheless, analysis of the dynamics observed when the robots interact with each other then revealed exactly such complex behaviours that are mathematically not possible within the dynamic range of the controller alone. This provides a very pithy demonstration that a neural system may behave differently when studied in isolation compared to its actual use case in interaction with other agents.

A difference between traditional computationalist approaches to artificial agents and present-day neuromorphic ones are that the former are based on classic computationalism. In such an approach, the perturbations that a human can bring into the system are more controlled and well-defined because they essentially reduce to known states (this also makes the systems brittle – they cannot deal as well with situations not known at design time – which is arguably one of the reasons for pushing towards more robust neuromorphic solutions) and traditional HCI approaches to the evaluation of the system are sufficient.

When the artificial agents are fundamentally neuromorphic, however, the interaction is arguably one between two dynamical systems that are able to perturb each other's dynamics in non-trivial ways. This is a strong reason to not ignore human end users while designing a system – the system's behaviour can potentially only be adequately characterised when it is in interaction with these users. At present, our evaluation metrics, whether it is for safety or other aspects such as explainability, do not fully capture this. More research is needed to understand how to best address this, but a viable starting point seems to exist within enactivist approaches to characterising natural cognition. It is worth noting that these approaches are amenable to studying social interactions between agents (Froese and Di Paolo, 2008; Froese and Paolo, 2010; Candadai et al., 2019). It is also worth not-

ing that this would complement, rather than replace, other approaches such as those reviewed above. There will still be a need to verify that a system is at least in principle functioning as intended. The point here is just that, once it is ready for interaction with human end users, these humans co-determine the behaviour of the system in non-trivial ways when the system is no longer simply built upon a computationalist paradigm, and the details of this are not yet fully understood.

4 Conclusions

In this brief note, we were concerned with the role that users play in interaction with autonomous technology, and how one could evaluate such advanced systems. Ensuring the safety of learning, perception and control in real world settings requires that the corresponding evaluations do not just reduce humans into the role of passive (in terms of system behaviour) end users but rather as part of the overall system. In the cognitive sciences, enactivism has a long history of studying agents in such terms and, even if one disagrees with it philosophically, it may be interesting to look towards it in order to better understand how to evaluate near future automated technology, whether vehicles or robots. This seems particularly pertinent if these are controlled using neuromorphic architecture because this does fundamentally create the kind of interactions – namely those between dynamical systems – that enactivism seeks to characterise.

Acknowledgements

ST is supported by the EC H2020 research project Dreams4Cars (no. 731593).

MR is supported by the Swedish Research Council project EXPLAIN, Evaluation of eXplainable Artificial Intelligence (VR 2018-03622).

References

Biran, O. and Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 8.

Blouw, P., Choo, X., Hunsberger, E., and Eliasmith, C. (2018). Benchmarking Keyword Spotting Efficiency on Neuromorphic Hardware. *arXiv:1812.01739 [cs, stat]*. arXiv: 1812.01739.

Boukhelifa, N., Bezerianos, A., and Lutton, E. (2018). Evaluation of Interactive Machine Learning Systems. *arXiv preprint arXiv:1801.07964*.

Breazeal, C. (2004). Social interactions in hri: the robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):181–186.

Candadai, M., Setzler, M., Izquierdo, E. J., and Froese, T. (2019). Embodied dyadic interaction increases complexity of neural dynamics: A minimal agent-based simulation model. *Frontiers in Psychology*, 10:540.

Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A. V., and Turrin, R. (2011). Looking for ?good? recommendations: A comparative evaluation of recommender systems. In *IFIP Conference on Human-Computer Interaction*, pages 152–168. Springer.

Da Lio, M., Mazzalai, A., Windridge, D., Thill, S., Svensson, H., Yksel, M., Gurney, K., Saroldi, A., Andreone, L., Anderson, S. R., and Heich, H. (2017). Exploiting dream-like simulation mechanisms to develop safer agents for automated driving: The dreams4cars eu research and innovation action. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of humanrobot interaction. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1480):679–704.

De Jaegher, H., Di Paolo, E., and Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, 14(10):441–447.

DeWolf, T., Stewart, T. C., Slotine, J.-J., and Eliasmith, C. (2016). A spiking neural model of adaptive arm control. *Proceedings of the Royal Society B: Biological Sciences*, 283(1843):20162134.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.

Dimara, E., Bezerianos, A., and Dragicevic, P. (2018). Conceptual and Methodological Issues in Evaluating Multidimensional Visualizations for Decision Support. *IEEE Trans. on Vis. & Comp. Graphics*, 24(1):749–759.

Froese, T. and Di Paolo, E. (2008). Stability of coordination requires mutuality of interaction in a model of embodied agents. In *From Animats to Animals 10, The 10th International Conference on the Simulation of Adaptive Behavior*, volume 5040, pages 52–61.

Froese, T., Gershenson, C., and Rosenblueth, D. A. (2013). The dynamically extended mind. In *2013 IEEE Congress on Evolutionary Computation*, pages 1419–1426.

Froese, T. and Paolo, E. A. D. (2010). Modelling social interaction as perceptual crossing: an investigation

into the dynamics of the interaction process. *Connection Science*, 22(1):43–68.

Gregor, S. and Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, pages 497–530.

Groom, V. and Nass, C. (2007). Can robots be teammates?: Benchmarks in humanrobot teams. *Interaction Studies*, 8(3):483–500.

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Kirsch, A. (2018). Explain to whom? Putting the User in the Center of Explainable AI. In *IJCAI-17 Workshop on Explainable AI (XAI)*.

Lagerstedt, E., Riveiro, M., and Thill, S. (2017). Agent autonomy and locus of responsibility for team situation awareness. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, pages 261–269, New York, NY, USA. ACM.

Lim, B. Y., Dey, A. K., and Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 2119–2128. ACM.

Maturana, H. R. and Varela, F. J. (1987). *The tree of knowledge: The biological roots of human understanding.* New Science Library/Shambhala Publications.

Mohseni, S., Zarei, N., and Ragan, E. D. (2018). A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839*.

Plebe, A., Da Lio, M., and Bortoluzzi, D. (2019). On reliable neural network sensorimotor control in autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–12.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proc. 22nd SIGKDD Int. Conf. on Know. Discovery and Data-Mining*, pages 1135–1144. ACM.

Springer, A. and Whittaker, S. (2018). " i had a solid theory before but it's falling apart": Polarizing effects of algorithmic transparency. *arXiv preprint arXiv:1811.02163*.

Talbot, J., Lee, B., Kapoor, A., and Tan, D. S. (2009). EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers. In *Proc. of the 27\textsuperscriptth Int. Conf. on Human Factors in Computing Systems (CHI'09)*, pages 1283–1292, New York, NY, USA. ACM. event-place: Boston, MA, USA.

Thill, S., Hemeren, P. E., and Nilsson, M. (2014). The apparent intelligence of a system as a factor in situation awareness. In *Proceedings of the 4th IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, pages 52 – 58.

Thill, S., Riveiro, M., Lagerstedt, E., Lebram, M., Hemeren, P., Habibovic, A., and Klingegrd, M. (2018). Driver adherence to recommendations from support systems improves if the systems explain why they are given: A simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 56:420–435.

Thill, S. and Ziemke, T. (2015). Interaction as a bridge between cognition and robotics. In *"Cognition as a bridge between robotics and interaction" workshop in conjunction with HRI2015*.

Yi, J. S., Kang, Y.-a., Stasko, J. T., and Jacko, J. A. (2008). Understanding and characterizing insights: how do people gain insights using information visualization? In *Proceedings of the 2008 conference on BEyond time and errors*, pages 1–6, New York, NY, USA. ACM. event-place: Florence, Italy.