Artificial Phenomenology for Human-Level Artificial Intelligence

Lorijn Zaadnoordijk[1] and Tarek R. Besold[2]

[1] Radboud University, Donders Institute for Brain, Cognition, and Behaviour, Nijmegen, The Netherlands
L.Zaadnoordijk@donders.ru.nl
[2] Alpha Health AI Lab, Telefonica Innovation Alpha, Barcelona, Spain
Tarek.Besold@telefonica.com

Abstract. For human cognizers, phenomenal experiences take up a central role in the daily interaction with the world. In this paper, we argue in favor of shifting phenomenal experiences into the focus of human-level AI (HLAI) research and development. Instead of aiming to make artificial systems feel in the same way humans do, we focus on the possibilities of engineering capacities that are functionally equivalent to phenomenal experiences. These capacities can provide a different quality of input, enabling a cognitive system to self-evaluate its state in the world more efficiently and with more generality than current methods allow. We ground our general argument using the example of the sense of agency. At the same time we reflect on the broader possibilities and benefits for artificial counterparts to human phenomenal experiences and provide suggestions regarding the implementation of functionally equivalent mechanisms.

Keywords: Human-Level Artificial Intelligence · Phenomenology · Sense of Agency.

1 Introduction

Phenomenal experiences are a defining element of many interactions with our surrounding world. While to us the presence of phenomenal qualities in our everyday cognition does not always deserve active attention, the disappearance of the experiential dimension would have far-reaching ramifications, for instance, for learning, social interaction, and ethical behavior. Phenomenology has, therefore, been a popular topic of theoretical and empirical investigation across different disciplines [3, 7, 19] but—bar a few laudable exceptions such as [17, 4]—has been widely ignored in AI. We argue in favor of shifting phenomenology also into the focus of human-level AI (HLAI) research and development. Phenomenal experiences provide a different quality of input to cognition as compared to non-phenomenal perception (i.e., abstract registration of stimuli from the environment). Among others, phenomenology can facilitate the self-evaluation of an artificial cognitive system's state in the world, facilitating learning about and interacting with the physical world and other agents.

2 Phenomenology in Human-Level Artificial Intelligence

HLAI aims at developing machines that can meaningfully be considered to be on par with humans in that they are similarly able to reason, to pursue and achieve goals, to perceive and respond to different types of stimuli from their environment, to process information, or to engage in scientific and creative activities. Our view on HLAI is functionalist: Any technologically realizable means of (re)creating human-level intelligence in an artificial system are considered valid.

One of the core challenges any kind of cognitive system needs to solve is how to best interact with the world (i.e., the environment in which it is situated) and the associated (self-)evaluation of its state in the world. For humans, at least two possible ways of solving these interconnected problems come to mind: one route draws upon high-level reasoning capacities, and another one relies on phenomenal experiences. The former route likely draws on a process requiring all of perception, representation, reasoning, and evaluation. Phenomenal experiences on the other hand often take over the function of providing immediate—and in comparison much more unmediated—access and evaluation, allowing to go from perception to evaluation via a route not involving high-level reasoning:

1. Perceive sensory input(s) $\{Y\}$.
2. Represent the perceived inputs: $R_2(Y)$.
3. Map from $R(\{Y\})$—and system-internal information S—to an evaluation of the experiential category, quality, and valence in terms of, e.g., pain or pleasure, weak or strong, attractive or aversive: $E(R(Y), S) \mapsto \{\{\mathsf{pain}, \mathsf{pleasure}, \ldots\} \times \{\mathsf{weak}, \mathsf{strong}, \ldots\} \times \{\mathsf{attractive}, \mathsf{aversive}, \ldots\}, \emptyset\}$.[3]

Comparing both approaches, three advantages of the second route involving phenomenal experiences can be explicated: (i) increased efficiency and tractability, (ii) reduced requirements regarding additional information, and (iii) increased generality. Mapping directly from perceptual representations to evaluations removes the reasoning process from representation to category label which otherwise is likely to involve the exploration of a significantly-sized state space or the execution of a lengthy chain of individual reasoning steps. Moreover, the successful performance of the high-level reasoning mechanism in many cases requires further knowledge, which might not be available to the cognizer at the relevant point in time. Phenomenal experiences, by contrast, are assumed to be mostly independent from a person's knowledge (although they might be influenced by prior experiences and familiarity with a percept). Finally, in 'standard' approaches the interface between system and environment is commonly conceived of in terms of evaluative functions taking two sets of input: A set of current system and world states, often together with representations of potential actions of the system, and a set of goals (i.e., desired system or world states). The function output is an evaluation of the system and world states relative to the systems goals. Generating these functions is far from trivial and hitherto lacks

[3] The codomain of the mapping includes $\emptyset$ to account for cases where perception does not yield a phenomenal experience as is the case, e.g., in subliminal priming.

a general answer or methodology. In most cases this hinders generalizability as evaluation functions have to be grounded in a certain domain or action space to be definable in a comprehensive way. Moreover, they rely on the presence (or absence) of certain defined domain elements or action possibilities which imposes further limitations regarding the generality of application domains.

3 Functional Equivalence in Artificial Phenomenology

We argue for engineering *artificial phenomenology* (i.e., a functional equivalent of phenomenal experiences) rather than human-like phenomenal experiences. Even if we knew how to reproduce human phenomenology in artificial systems, due to a lack of kinship between AI/robotic systems and humans assuming similarity of the phenomenal experience a priori is unwarranted: It might well be the case that the precise phenomenal qualities may be an epiphenomenon resulting from the particular forms of representation and/or processing in humans [6]. Still, we believe that identity of phenomenal experiences is not required, but that a functional equivalent on the side of the machine suffices for the purposes of creating HLAI. The challenge, thus, becomes one of engineering a capacity fulfilling the same functions as phenomenal experiences do within cognitive processes, but remains agnostic regarding the actual qualitative dimension.

In considering ways of implementing artificial phenomenology, we take a representationalist approach [5] as often applied both to cognitive capacities as well as phenomenal experiences. Representationalist accounts of phenomenology posit that experiential states are characterized by the representational content [3]. Representationalism offers a natural interface to approaches in HLAI, which build upon the computational cognition maxim (i.e., assuming that computation is in principle capable of giving rise to mental capacities) and, therefore, among others introduce representations as important part of computation [14, 16].

4 Implementing Phenomenology: the Sense of Agency

In typically-developed human adults the "sense of agency" (i.e., the feeling of causing one's actions and their consequences [10]), contributes to important aspects of cognition, such as learning through intervention [11], social and moral interaction [2], and self-other distinction [19]. At least two different phenomena are considered under the banner of the "sense of agency": the "judgment of agency" and the "feeling of agency" [18]. In the case of the judgment, a reasoning step gives rise to the assumed status as agent in the world—considering oneself as agent provides the best explanation for the observations from the environment, thus agency is assumed in a post-hoc fashion. This results in a belief state ascribing agency to the reasoner. In the case of the feeling of agency, agency is not directly perceived nor concluded as outcome of an active reasoning process but is experienced as a phenomenal quality based on a representation of what the world is like. In contrast to the judgment, the feeling of agency is, thus, more akin to a perceptual state than to a belief state.

From an HLAI perspective, both concepts pose different challenges when considering an implementation. The judgment of agency requires a reasoning process determining oneself as the most likely cause for the observed changes in the state of the environment. Implementing this reasoning in a cognitive system returns to several facets of the Frame Problem [9]. If demanding the judgment of agency to be infallible, the system must be able to rule out all possible alternative causes (observed and unobserved) for the respective change in the world. Alas, already deciding which aspects of the perceptual input are relevant for performing the judgment of agency carries the danger of computational intractability, as does the subsequent reasoning process. Luckily, infallibility imposes an unreasonable standard not least because also humans can err when being asked to judge their agency in settings where an immediate observation is not possible [20]. In practical terms, implementing the judgment of agency becomes equivalent to a form of inference to the best explanation—and, thus, to abductive reasoning [13, 8]: The system must decide if a change in its environment is most likely due to its own actions, making it an agent within the corresponding situation (or not).

The feeling of agency as perceptual state is often thought to arise from a comparison between the predicted state of the world following one's action on the one hand, and the observed state of the world on the other hand [1]. Motivated by its contribution to other cognitive capacities, several groups of researchers have engaged with the question of how to equip artificial systems with a SoA or related capacities [15, 12]. A commonality of these and similar projects is their primary focus on contingency detection. However, while contingency detection plays a major role in human SoA, by itself it is not sufficient [21]. Instead, it likely is the case that the detection of a sensorimotor contingency serves as a cue for the causal inference that the action was caused by oneself, the output of which is a mental causal representation that in part characterizes the SoA. Returning to the context of artificial cognitive systems, obtaining the corresponding evaluations also necessitates one further step beyond the detection of the contingency between predicted and observed world state. Still, while inferential in nature, this step does not have to involve forms of complex high-level reasoning as would be the case for the judgment of agency. It could be carried out following the general pattern for phenomenal experiences laid out above: Provided with the perceived world state as sensory input, and the predicted world state within the system-internal information at the current point in time, the detection of an equality relation between both causes a mapping to 'sense of agency' as experiential category. This, of course, unavoidably triggers the question for the genesis of the required mapping function. Different approaches are imaginable, including *a priori* hardcoding by the system architect, learning from observed statistical regularities, or via an explicit 'teaching' effort through the designer or the user.

Generally, the challenge in engineering a functional equivalent of the human feeling of agency resides in leaving out the actual qualitative dimension of human phenomenal experiences (cf. the corresponding discussion in Section 3) without also stripping away the benefits of having phenomenal experiences. A possible solution is a direct mapping of certain sensory ranges combined with a snapshot of the internal system state onto immediate "phenomenal values". Given the system state at one particular point in time, certain sensory inputs ought to give rise to artificial phenomenology. Artificial counterparts of phenomenal experiences and their rich qualitative properties can be defined as immediate mappings from the output ranges of the available sensors of the system, combined with specific information regarding the internal state of the system. At this point, the important property is the finite and known range of both, the sensors and the internal representational mechanisms of the system. By cutting out the reasoning step, the phenomenally-inspired approach neither requires an exhaustive enumeration and interpretation (and, thus, in practice a restriction) of the space of possible percepts and their representations, nor does it involve an often times computationally costly evaluation of the current system and world state relative to any goal state(s). Reducing the relevant information to the percept representations together with system-internal properties, and applying a direct mapping to qualitative categories with associated evaluation values therefore increases the tractability of the computational process and the generality of the approach. The output values can then serve as direct functional counterparts of human phenomenal experiences, for example triggering evasive reactions if "pain" is encountered or providing positive reward and consequently motivation to continue an action if "pleasure" arises.

5 Conclusions

We have argued that imbuing HLAI systems with capacities paralleling the role of phenomenal experiences in humans facilitates learning and acting in the world in a more general and tractable way than currently possible. Returning to our comparison in Section 2 between a process involving perception, representation, reasoning and evaluation versus the shorter perception-representation-evaluation cycle of artificial phenomenology, the latter promises to enable the system to self-evaluate its state in the world without the use of knowledge-rich, domain-specific evaluation functions or intractable reasoning processes. This could in turn facilitate learning and acting in the world in terms of assigning actions based on their predicted outcomes and assessing actual action outcomes.

In terms of applications, beyond the already mentioned obvious advantages regarding the progress towards creating HLAI as a research endeavor, artificial phenomenology promises to unlock a new qualitative dimension in human-computer interaction (HCI) settings. Artificial phenomenology would greatly contribute to system behaviour closer resembling human agents, as well as to complex user-modelling capacities providing more immediate—and likely generally better-informed—accounts of a user's cognitive state(s) as basis of interaction and collaboration. As such, several aspects motivate the need for artificial phenomenology and, therefore, the need for research into the possibilities. In this paper, we have outlined a starting position for this enterprise.

References

1. Blakemore, S.J., Wolpert, D.M., Frith, C.D.: Central cancellation of self-produced tickle sensation. Nature Neuroscience **1**(7), 635 (1998)
2. Caspar, E.A., Cleeremans, A., Haggard, P.: Only giving orders? An experimental study of the sense of agency when giving or receiving commands. PloS ONE **13**(9), e0204027 (2018)
3. Chalmers, D.J.: The representational character of experience. The Future for Philosophy pp. 153–181 (2004)
4. Chella, A., Manzotti, R.: Artificial consciousness. In: Cutsuridis, V., Hussain, A., Taylor, J.G. (eds.) Perception-Action Cycle: Models, Architectures, and Hardware, pp. 637–671. Springer New York, New York, NY (2011)
5. Cummins, R.: Meaning and Mental Representation. MIT Press (1989)
6. Dehaene, S., Lau, H., Kouider, S.: Response to commentaries on what is consciousness, and could machines have it?. Science **359**(6374), 400–402 (2018)
7. Dehaene, S., Naccache, L.: Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition **79**(1-2), 1–37 (2001)
8. Denecker, M., Kakas, A.: Abduction in logic programming. In: Kakas, A.C., Sadri, F. (eds.) Computational Logic: Logic Programming and Beyond: Essays in Honour of Robert A. Kowalski Part I, pp. 402–436. Springer, Berlin/Heidelberg (2002)
9. Dennett, D.: The frame problem of ai. Philosophy of Psychology: Contemporary Readings **433**, 67–83 (2006)
10. Haggard, P., Chambon, V.: Sense of agency. Current Biology **22**(10), R390–R392 (2012)
11. Lagnado, D.A., Sloman, S.: Learning causal structure. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 24 (2002)
12. Lara, B., Hafner, V.V., Ritter, C.N., Schillaci, G.: Body representations for robot ego-noise modelling and prediction. towards the development of a sense of agency in artificial agents. In: Proceedings of the Artificial Life Conference 2016 13. pp. 390–397. MIT Press (2016)
13. Mooney, R.J.: Integrating abduction and induction in machine learning. In: Flach, P.A., Kakas, A.C. (eds.) Abduction and Induction: Essays on their Relation and Integration, pp. 181–191. Springer Netherlands, Dordrecht (2000)
14. O'Brien, G., Opie, J.: The role of representation in computation. Cognitive Processing **10**(1), 53–62 (2009)
15. Pitti, A., Mori, H., Kouzuma, S., Kuniyoshi, Y.: Contingency perception and agency measure in visuo-motor spiking neural networks. IEEE Transactions on Autonomous Mental Development **1**(1), 86–97 (2009)
16. Rescorla, M.: Computational modeling of the mind: what role for mental representation? Wiley Interdisciplinary Reviews: Cognitive Science **6**(1), 65–73 (2015)
17. Sloman, A., Chrisley, R.: Virtual machines and consciousness. Journal of consciousness studies **10**(4-5), 133–172 (2003)
18. Synofzik, M., Vosgerau, G., Newen, A.: Beyond the comparator model: A multifactorial two-step account of agency. Consciousness and Cognition **17**(1), 219–239 (2008)
19. Tsakiris, M., Schütz-Bosbach, S., Gallagher, S.: On agency and body-ownership: Phenomenological and neurocognitive reflections. Consciousness and Cognition **16**(3), 645–660 (2007)
20. Wegner, D.M.: The Illusion of Conscious Will. Bradford Books/MIT Press (2002)
21. Zaadnoordijk, L., Besold, T.R., Hunnius, S.: A match does not make a sense: On the sufficiency of the comparator model for explaining the sense of agency (submitted)