Bayesian ghosts in a machine?

Martin Biehl[1*] and Nathaniel Virgo[2]

[1]Cross Labs, Cross Compass, Tokyo 104-0045, Japan
[*]martin.biehl@cross-compass.com
[2]Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo 152-8550, Japan

Abstract

We informally summarize our recent work on Bayesian reasoners and agents. We also briefly sketch its relation to an existing enactive definition of agents.

Introduction

We take the perspective that the world is composed of interacting systems. But is there something special about that subset of systems that we call 'living'?

One feature that has been proposed to help in distinguishing living from non-living systems is agency. However, there is no established consensus on when a system constitutes an agent.

In Virgo et al. (2021) we proposed conditions under which an open dynamical system can be called a *Bayesian reasoner* i.e. can be interpreted as having a model and maintaining subjective probabilistic beliefs over the model's hidden variables. Building on this we extended these conditions in Biehl and Virgo (2023) to capture when an open dynamical system can be called a *Bayesian agent*, i.e. can be interpreted as a Bayesian reasoner that additionally acts purposefully to achieve a goal.

These conditions provide new conceptual approaches to understanding agency that involve notions like beliefs, models and reasoning. This is in contrast (but not necessarily in contradiction) to other approaches like that of Barandiaran et al. (2009) that avoid such notions.

On a very high level the idea behind these conditions is to specify exactly when the dynamics of the system *coincide* or *are consistent* with those we expect from a system that uses a model, updates its beliefs about the hidden variables in the model, and maybe even takes actions according to those beliefs in order to achieve a goal.

The dynamical systems considered are *open* which means they usually have inputs and outputs. For this reason they are also referred to as *(state) machines*. As we will see, the models, beliefs, and goals, need not explicitly exist anywhere "inside" the machines. It is the existence of a relation between the dynamics of the machine's state and the dynamics of beliefs in belief space that allows us to interpret the system as having a model and acting according to its beliefs to achieve a goal. Metaphorically and in reference to the conference motto, "ghost in the machine", if we consider what happens in state space as what is "mechanical" then, the dynamics of the beliefs imposed by model and goal may be seen as more ghost-like, although they are no less mathematically real.

Bayesian reasoners

Consider a discrete time open dynamical system or machine with constant (or equivalently no) output. Such a system is given by a function[1] $\mu : I \times M \to M$ taking an input $i \in I$ and internal state $m \in M$ to a new internal state $m' = \mu(i, m)$. Inputs must come from somewhere e.g. they could be produced as the outputs of another machine or just taken from a long list of inputs one after the other. It does not matter here. Naively, what we want to know is whether this machine "contains" a model of the origin of its inputs and "has" beliefs about hidden variables in this model. As mentioned we would like to answer this question solely in terms of properties of the machine / dynamical system.

This can probably be done more generally for any well defined notion of beliefs and belief updates (using a model or not) but in Virgo et al. (2021) we employ the notion of beliefs as probability distributions that together with a probabilistic model have a well known belief update rule i.e. Bayes rule. We now briefly sketch the idea.

Recall that Bayes' rule takes as input a model of how some (possibly dynamic) hidden variable produces observations, a (prior) belief over the hidden variable's value, and an observation, and tell us as its output what the (posterior / updated) belief over the hidden variable should be after the observation has been made.

To find out whether μ is a Bayesian reasoner we guess

[1]Stochastic versions of this are possible but the deterministic one is easier and we believe more instructive since it highlights that deterministic systems can have interpretations in terms of *probabilistic* beliefs.

- a model[2] $\kappa : H \to P(H \times I)$
- an *interpretation map*, which is a function $\psi : M \to PH$.

The map $\psi : M \to PH$ associates to each internal state $m \in M$ of the open dynamical system a probability distribution (more intuitively, a probabilistic belief) $\psi(m)$ over the hidden variable $h \in H$ of the model κ. Note that this means that whenever the state $m \in M$ of the system changes in response to an input $i \in I$ we obtain not only a new state $m' = \mu(i, m)$ but we also get the new belief $\psi(m') = \psi(\mu(i, m))$ that is associated to it. Indeed, it is possible that this new belief coincides with the Bayesian posterior with respect to the prior $\psi(m)$ associated to the original state, model κ and and observation i. In that case the guess is successful: the interpretation map does indeed give us a consistent interpretation of the system as a Bayesian reasoner. In Virgo et al. (2021) we give an explicit equation that expresses this consistency.

This means the map ψ extracts from the machine μ the dynamics of Bayesian belief updating with respect to model κ and observations equal to its inputs. In this sense, it's justified to interpret the machine μ as having the model κ and maintaining Bayesian beliefs about its hidden variables.

Such interpretations are not unique, however: the same machine may be interpreted as a Bayesian reasoner in many different ways, i.e. there might be many different maps ψ and models κ that form consistent interpretations in the sense described above.

Bayesian agents

Consider now the case where a machine μ like above also has an output function $\omega : M \to O$ that produces an output $\omega(m)$ for each internal state $m \in M$. We may want to know if we can interpret this machine as trying to solve a problem or achieve a goal.

Again, this can probably be done more generally but in Biehl and Virgo (2023) we provide concrete conditions for the class of partially observable Markov decision problems (POMDPs). This choice also directly extends the Bayesian reasoner above. We now also briefly sketch the idea.

To find out whether μ and ω form a Bayesian agent we guess

- a model $\kappa : H \times O \to P(H \times I)$
- a reward function $r : H \times O \to \mathbb{R}$
- an interpretation map $\psi : M \to PH$.

The model κ now also takes an output $o \in O$ of our machine as an argument and thus takes into account the influence of this output on the hidden variable $h \in H$. The reward function encodes the goal, which is the maximization of the expected discounted cumulative reward. The map ψ plays the same role as above: we require that ψ maps the states to beliefs about the hidden variable in such a way that on the level of associated beliefs priors get updated to Bayesian posteriors with respect to model κ, input $i \in I$, and output $o = \omega(m) \in O$.

However, for a successful guess we now need another coincidence. The optimal actions for achieving the goal implied by κ and r must coincide with the outputs of the machine. Since the problem is partially observable the optimal action can be expressed as a function of probabilistic beliefs over the hidden variable. So in turns out that the output $\omega(m)$ for state m must coincide with the action that is optimal for the belief $\psi(m)$ that is associated to this state m.

If both these conditions are satisfied then ψ again extracts from the state dynamics of μ the belief dynamics with respect to the model κ. Additionally ω specifies just those outputs that would be the optimal actions for achieving the goal defined by r under the beliefs extracted by ψ. In this sense, the dynamical system can be interpreted as having a model, maintaining Bayesian beliefs, and acting purposefully according to those beliefs in order to achieve a goal.

Discussion

A question that often arises with respect to definitions of agents is in how far they require an observer to distinguish them from their environment or whether the agents "actively define themselves" (Di Paolo et al., 2017, p.112). Clearly, open dynamical systems do not actively define themselves so the notion of a Bayesian agent is incomplete in this sense.

One may also wonder whether an observer is required in another sense. We have given conditions under which a machine can be interpreted as a Bayesian reasoner or an agent, but do we need an observer to perform such an interpretation? On the one hand, whether such interpretations exist or not is an observer-independent mathematical fact, but on the other hand they are not unique, so perhaps we need an observer to choose one. (cf. Dennett's intentional stance.)

Bayesian interpretations may be useful to formalize what has been termed "interactional asymmetry" between agent and its environment. Di Paolo et al. (2017) write "Intuitively it would seem right to equate [interactional asymmetry] with the idea of agents being the "cause" of certain events." Assume we have two interacting machines. If one has an interpretation as a Bayesian agent its outputs are optimal actions with respect to its beliefs to achieve some goal. This may justify referring to its outputs and those consequences predicted by the agent's model as "caused" by the agent, i.e. as caused by the ghost and not just the machine.

Acknowledgements

This project was made possible through the support of Grant 62229 from the John Templeton Foundation. The opinions

[2]We write PX for the set of probability distributions over the set X. Specifically $P(H \times I)$ is the set of probability distributions over pairs $(h, i) \in H \times I$.

expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. The work was also supported by a grant from GoodAI.

References

Barandiaran, X. E., Paolo, E. D., and Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5):367–386.

Biehl, M. and Virgo, N. (2023). Interpreting systems as solving POMDPs: A step towards a formal understanding of agency. In Buckley, C. L., Cialfi, D., Lanillos, P., Ramstead, M., Sajid, N., Shimazaki, H., and Verbelen, T., editors, *Active Inference. IWAI 2022. Communications in Computer and Information Science*, pages 16–31. Springer.

Di Paolo, E., Buhrmann, T., and Barandiaran, X. (2017). *Sensorimotor Life: An enactive proposal*. Oxford University Press.

Virgo, N., Biehl, M., and McGregor, S. (2021). Interpreting dynamical systems as bayesian reasoners. In Kamp, M. et al., editors, *International Workshops of ECML PKDD 2021*, pages 726–762. Springer.