DOI: http://dx.doi.org/10.7551/978-0-262-33027-5-ch089

Apparent actions and apparent goal-directedness

Martin Biehl and Daniel Polani

University of Hertfordshire, Hatfield, UK

m.biehl@herts.ac.uk

Introduction In human history countless phenomena have been (wrongly) attributed to *agents*. For instance, now science believes there are no gods (agents) of lightning, thunder and wind behind the associated phenomena.

In physics (assuming quantum decoherence) the universe is modelled as a state space with a dynamical law that determines everything that happens within it. This however, is incompatible with most notions of agency (cf. Barandiaran et al., 2009) which require *actions*: For an agent candidate to have actions it must be able to "make something happen" as opposed to only "have things happen to it".

Here we ask which single sequences of partial observations may *appear* to contain agency to a *passive observer* who has its own memory. For this we define measures of *apparent actions* and *apparent goal-directedness*. Goal-directedness is another feature commonly attributed to agents. We here ignore whatever causes the appearances and the concept of individuality of agents.

Apparent actions We assume that a passive observer perceives a sequence S of sensor values $(s_1, ..., s_T)$ with $T \in \mathbb{N}^+$. At each instance t of the sequence, the observer has a memory or knowledge state m_t which gives us a second sequence $M = (m_1, ..., m_T)$. The memory state of the agent might contain models of the observations or not. In the course of a sequence S there is an *apparent action for observer* M if for $r, t \in \{1, ..., T\}$ we have $(s_r, m_r) = (s_t, m_t)$ and $(s_{r+1}, m_{r+1}) \neq (s_{t+1}, m_{t+1})$.

The intuition is that, since the same observation $s_r = s_t$ at different times r, t is followed by different observations and the observer's states $m_r = m_t$ do not indicate / predict the difference, the observer suspects a hidden mechanism causing the difference. We assume that the observer interprets all signs of hidden mechanisms as (apparent) actions.

Using the empirical distribution $p_{S,M}(s', s, m) = \frac{1}{T}\sum_{t=0}^{T} \delta_{s's_{t+1}}\delta_{ss_t}\delta_{mm_t}$ (δ_{xy} is Kronecker's delta) we can quantify the extent of apparent actions along the sequences as the conditional entropy $\mathrm{H}(S'|S, M)$.

Apparent goal-directedness Our observer attributes the complete sequence S to be the result of an agent's strategy. Note that even if there is no apparent action along a subsequence this could be due to the agent trying to avoid detection. The idea is that any directedness reveals itself as some pattern within the sequence and any pattern in the sequence will increase the compressibility of the sequence. So we here define *apparent goal-directedness of the observed sequence* S as its compressibility. Using a common compression algorithm (e.g. gzip) Z we can estimate compressibility as $l(S) - Z(S)$ where $l(S)$ is the binary length of all the data in the observed sequence and $Z(S)$ the binary length of the compressed data. Note that an adversary's goal-directedness can remain undetected only if S is completely random.

Examples Requiring both apparent action and apparent goal-directedness leads to the following classifications: 1.) A Brownian particle exhibits apparent actions but very low apparent goal-directedness. 2.) A ball thrown through the air exhibits no apparent actions if $m_t = s_{t-1}$ i.e. if memory contains the previous observations. Together the s_{t-1} and s_t are enough to get linear momentum and position of the ball which together determine its trajectory. Apparent goal-directedness is high as the equations of motion compress the flight path. 3.) A thief trying to guess a safe combination exhibits apparent actions as every time a new combination is tried out it starts in the same initial position. It also exhibits some goal-directedness, as it never tries the same combination twice, which is a pattern.

Conclusion The apparent notions identify agency in example systems and take into account capabilities of the observer. To fool the observer a (visible) adversary has to *a*) be predictable to the observer (no apparent action) or *b*) rely on randomness (luck) to achieve its goal (no goal-directedness).

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.

Martin Biehl, Daniel Polani (2015) Apparent actions and apparent goal-directedness. Proceedings of the European Conference on Artificial Life 2015, p. 511