THE LION, THE CHIMP AND THE BANANAS

J E Tardy
Systems Analyst
Sysjet inc.
jetardy@**sysjet.com**

Humans discard as "mechanical", a behaviour they can fully predict and as "meaningless", an output they view as random. However, a behaviour they perceive as ***both intentional and unpredictable*** *fascinates them.*

*Software applications are designed to function predictably or, in some cases such as game scenarios, their output is varied randomly. A system, to be perceived as consciously intelligent, should behave differently. Its behaviour should be purposeful without being predictable. Expressed in terms of model-predictive control, it should maintain its users in a state of "****Perceived Unpredictable Optimality****". A game scenario involving a lion, a chimpanzee and some bananas illustrates this concept and provides a template for its implementation.*

Date: Dartmouth NS – 2018.09.28
Keywords: Synthetic Consciousness, Digital Consciousness, Artificial Intelligence, Systems analysis, Game Theory, AI.

PERCEIVED UNPREDICTABLE OPTIMALITY

A behaviour that is entirely predictable indicates that the entity generating it does not take into account the effect this predictability has on others. This deters from the perception, on the part of a user, that the entity is aware of its environment. A behaviour that a user perceives as random will be unpredictable but will also be discarded as a simple randomizing mechanism.

On the other hand, a behaviour that is perceived as goal-oriented but remains unpredictable has a fascinating effect. If a user perceives that a behaviour is goal-oriented but is nonetheless unable to predict it, he will tend to interpret this **unpredictable optimality** as both mysterious and complex. He will also attribute a higher degree of "awareness" to the entity that generates this perception.

If, in addition, a user perceives that the entity is taking his own cognitive perceptions into consideration to generate a behaviour that is **intentionally unpredictable** by him, this will have a powerful effect and generate, in the user, a sensation he is interacting with an entity that has a degree of **self-awareness**.

I refer to these perceptions, respectively, as **Perceived Unpredictable Optimality** and **Perceived Intentional Unpredictable Optimality**.

> **Perceived Unpredictable Optimality** is a user-state that is generated when a user detects the presence of an intentional pattern but remains unable to predict its future occurrences.
>
> **Perceived Intentional Unpredictable Optimality (PIOU)** occurs when a user concludes that the unpredictable optimality he perceives is determined by an internal model, generated within the entity, of the user's own predictive capabilities.

A system that generates a PIOU will have a powerful effect on its user. It may be perceived as a conscious intelligence regardless of its actual cognitive capabilities.

USER PERCEPTIONS

When a user observes the behaviour of a system, he can interpret it in one of **five ways**:

1. The behaviour follows a set pattern that is identically repeated in identical situations and thus **predictable**
2. The behaviour is generated by an optimizing model-predictive process that he understands and, thus, **can predict**.
3. The user considers the behaviour to be a **randomly generated choice** from available alternatives.
4. The user perceives the presence of a goal oriented pattern in the behaviour but is unable to generate predictive interpretations of it.

5. The user believes the system's behaviour is derived from a strategy that is **intentionally unpredictable** by taking his own cognitive capabilities into account to maintain the behaviour beyond predictability.

In the first case, the system behaviour is mechanically repeated and fully predictable. In the second case, if a goal is known to the user, then a strategy to optimize results will constantly select the best alternative and, thus, generate predictable output. In the third case, a randomly produced behaviour is unpredictable but the mechanism that generates it is simple and also, in this sense, "predictable".

On the other hand, a system whose behaviour maintains the user in state four achieves **Perceived Unpredictable Optimality**. If, in addition, the user also believes that the system's behaviour is conditioned by his own cognitive interpretation of the situation into account, he is in state five: **Perceived Intentional Unpredictable Optimality**.

SIMPLE TECHNIQUES

A system behaviour that exhibits unpredictable optimality has an interesting feature: it is very similar to an imperfect or suboptimal optimization process. It can be very difficult to distinguish whether a suboptimal choice is a simple flaw or part of a larger, misunderstood, strategy. Consequently, humans will often interpret a behaviour they can't predict as more complex than it is.

This similarity between suboptimal behaviour and intentional obfuscation together with the human propensity to attribute complex causes to misunderstood events makes it possible to generate the perception of unpredictable optimality with simple techniques.

> **Complex cognition is not essential to create the perception of complex cognition.**

When the original behaviour of a system is generated by model-predictive optimization, applying some simple techniques to modify the optimal choice will generate **Perceived Unpredictable Optimality**. For example:

- Randomly trigger occasional deviations from the optimal selection.
- Alternate the optimal behaviour selection process between different predictive models.
- Occasionally "flip" the behaviour by inverting the measure function used in the optimization.

- Use styling modifications to degrade the clarity of the output and make it ambiguous.

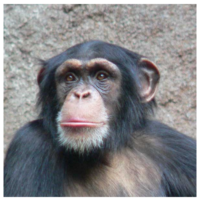
On a more advanced level, a system could also run a separate predictive modeling of its own behaviour and modify this behaviour whenever the match between actual and predicted output reaches a certain level of correctness.

These techniques retain the goal-oriented behaviour. They also mimic an intention to be unpredictable without actually modelling a user's cognitive perceptions. Many users would interpret the output generated by these simple techniques as emanating from complex cognitive processes.

THE LION, CHIMP, BANANA SCENARIO

The following game scenario models the production of Unpredictable Optimality and provides a template that can be used by learning systems to generate it. It also describes one of the elements that make Unpredictable Optimality fascinating: **recursive modelling**. The lion, the chimp and the bananas scenario defines a game-like situation where **recursively generating Unpredictable Optimality** is essential to win.

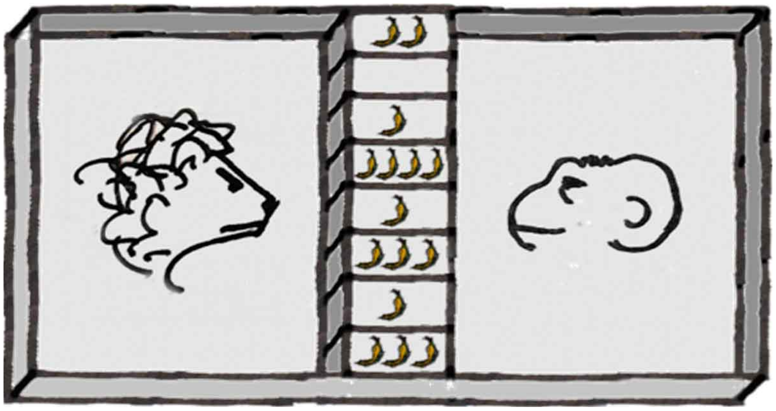
The game is played out in an imaginary zoo. This zoo has three sections, left, middle and right. The left section is a pen occupied by a lion. The right section is a yard where some chimpanzees reside. The middle section consists of a number of separate rooms (half a dozen or more). Each room has two doors, one opening to the lion's pen and the other to the chimp yard. However, the chimp side doors are too small for the lion to go through so both the lion and the chimps can go in the rooms but the chimps are safe in their yard.



A chimpanzee must eat a certain number of bananas to survive and the lion needs to catch and eat some chimp to stay alive.

Every morning the zookeeper goes through the middle section and randomly places a number of bananas in each room. For example, he may put three bananas in one room, one in another, no bananas in a third, and so on. The numbers of bananas differ each day but, each morning, both the lion and the chimp can "see" how many bananas each room contains

Once the zookeeper has placed all the bananas in the rooms, he opens the lion side doors. The lion chooses one of the rooms, enters it and hides, ready to pounce on the chimp. The zookeeper then opens the chimp-side doors.



The chimp can only choose one room each day and gets only the bananas it contains. If he wants to have bananas that day, he must choose one room, enter it and get those bananas. If the lion was hiding in that room then, arghh!, the lion eats the chimp. If the lion was not in the room, the chimp stays alive and has some bananas to survive.

Each afternoon, the zookeeper visits the lion in his pen and the chimp in his yard to tell them what happened in the morning. For example:

- *"Hello Mr. Lion, while you were hiding in room number 5, the chimp got the bananas in room number 3";*
- *"Hello Mr. Chimp, while you were getting the bananas in room number 3, the lion was waiting for you in room number 5".*

The next morning the process starts over.

At first, neither animal is very smart and they simply choose their rooms randomly. Later on, the potassium in the bananas makes the chimp smarter. He abandons the random selection approach and decides to enter the room that contains the most bananas to have more food. However, this optimal banana-eating behaviour does not take the lion's cognitive modelling into account and is very predictable.

Eventually, the lion, who was also choosing rooms randomly, becomes frustrated and that makes him smarter. He devises a cognitive representation of the situation that includes an internal "model avatar" of the chimp. In this model the chimp avatar wants to have as many bananas as he can. He uses this as a predictive model of the chimp's behaviour to determine his own choice: "*Choose the room with the most bananas*".

Luckily, the chimp's brain is now so full of banana vitamins and he has become even more intelligent. He develops an internal representation of the situation that also includes a "lion avatar" but, in addition, an avatar of himself. In this model, the chimp avatar chooses the room with the most bananas but, gasp, the lion avatar knows this and is waiting for him in that room. In other words, the chimp begins to recursively model the game in the sense that his cognitive representation now includes a "sub" representation of himself inside the lion avatar's internal representation of the situation.

However, the chimp still wants as many bananas as possible. He decides to avoid the optimal choice from now on and select the room with second most bananas.

After a few days without chimp meat, the lion, who was waiting in the room with the most bananas, is not only frustrated, he is hungry; and that makes him even smarter.

So he thinks even harder and adds another recursive layer to his internal model of the situation: The chimp's internal representation of the lion's internal representation of the chimp's internal representation of the situation. He realizes that the chimp knows that the lion knows that he wants to maximize bananas and so, avoids the room with the most bananas and chooses the second best banana alternative.

Of course, as the lion decides to hide in the room with the second most bananas, the chimp has figured out that the lion knows that he knows that the lion knows… and modifies his behaviour taking into account his enhanced understanding of the lion's internal model of the situation.

INTERPRETATION

The lion chimp banana interaction can be transposed to an UI interaction between a user and an application (app). In this transposition, the user is a lion and the app, a chimp. The user "pounces and eats" the app whenever it determines it is "*just a dumb program*". As the user interacts with the application, he "*lies in wait*" by mentally forecasting what the app will do next. This internal prediction is like choosing a room in the zoo. The user is then ready to "*pounce on the app*" by concluding that its behaviour is nothing more than the predictable result of a mechanical optimization process or is a randomly selected outcome that also makes no sense. In either case, the app is figuratively devoured and the user further feeds his chubby self-image.

Consequently, to "survive" its interactions with a user, an app, designed to be perceived as consciously intelligent, must constantly modify its behaviour to avoid predictability while also exhibiting a pattern that can be perceived, by the user, as intentional.

DESIGN NOTES

As the lion and the chimp get smarter, they will first adopt behaviours based on increasingly recursive models of the situation and of each other. However, in my view, as they develop ever more complex representations, the relative advantage of cognitive modeling will diminish.

At a certain point, advanced cognitive modeling itself, given the available information, no longer provides an advantage and may even become a deterrent. Eventually, an optimally unpredictable output is virtually identical with a randomly degraded output. This may be one reason why humans readily attribute the generation of unpredictable outcomes to complex cognitive processes.

> **Unpredictable optimality and limited randomization can be indistinguishable.**

In terms of design, **the simpler strategy of partially degrading an optimal output** by using techniques such as those outlined in the preceding section may be sufficient to generate Perceived Unpredictable Optimality.

A virtual instance of the Lion, Chimp, Banana interaction can be implemented in a virtual space and run iteratively using learning algorithms to generate strategies that are both unpredictable and effective with respect to a given objective. These strategies could then be **transposed to other situations** requiring unpredictable optimality.

Interestingly, behaving with **unpredictable optimality** in a way that is also perceived as such does not only depend on a system's internal modeling of the situation. It must also take into account the user's cognitive limits. A user that does not detect a complex but purposeful behaviour will perceive it as simply random. Based on a user's feedback, a system may have to "*dumb down*" or **simplify its behaviour to make sure the user perceives the pattern**.

The state of **Perceived Unpredictable Optimality** is primarily generated by the behaviour of a system. However, if the interaction with the user also includes communicated messages, these can be used to further enhance that state. Messages such as "*You probably expect your friends to come over*" or "*You are too predictable*" suggest that the system maintains an internal model of the user and has the capability to generate Perceived Intentional Unpredictable Optimality. As in the case of the behaviour itself, such messages can be effective whether the cognitive capability is present or not.

CONCLUSION

Humans disregard highly predictable behaviour as mechanical and purely random output as meaningless. However, humans are fascinated by behaviour they perceive as both intentional and unpredictable and tend to attribute a higher degree of awareness to the entity that generates it.

A system designed to be viewed as “consciously intelligent” must avoid behaviour that is perceived as either mechanical or meaningless. Instead, it must generate, in its users, a state of **Perceived Unpredictable Optimality**.

The game scenario of the Lion, the Chimp and the Bananas effectively models Perceived Unpredictable Optimality as well as recursive modeling. It also provides a template to automatically generate them.

Simple randomization and communication techniques can also produce **Perceived Unpredictable Optimality** without necessitating complex cognitive modeling.

JE Tardy

Note

The content of this article is adapted from ***Annex 8*** *of* ***The Meca Sapiens Blueprint****, a complete system architecture to implement digital consciousness with standard techniques and on conventional equipment*

The Meca Sapiens *Blueprint is available at* ***Glasstree Academic Publishing*** *and through* ***sysjet.com****.*