Consciousness and Causal Paradox of Emergent Systems

Jan Romportl

Department of Cybernetics,
University of West Bohemia in Pilsen
Univerzitní 8, Plzeň, Czech Republic
rompi@kky.zcu.cz

Abstract. This paper understands Golem as an artificial conscious being and the problem of Golem construction is seen as establishing appropriate causal relations between different causal domains – namely the causal domain of neural systems (either human brains or artificial neural networks) and the causal domain of mental processes (mental or psychological states and perhaps the consciousness). The paper further introduces and argues for the formal model of such causal domains (and their inter-domain relation) based on Vopěnka's Alternative Set Theory (AST). This model shows a very interesting result, which can be called "an inter-domain causal paradox": each phenomenon from the mental domain must be principally caused by all phenomena from the neural domain. The link between the emergentist paradigm and this ambiguous inter-domain causation is shown. This causal paradox is interpreted the way that vagueness (or "fundamental agnosticism") is a constitutive element and vital prerequisite for the consciousness existence. The paper concludes with doubts concerning the possibility of Golem creation.

1 Introduction

The topic of creation of an "artificial" conscious being (further in the text I will call it/him/her "Golem") is probably the heritage of our Judaic-Christian religious tradition together with the God's act of creating a human being (almost procedurally described in the Bible) and such as it is, it has been continually fascinating the mankind since then.

The emergentist paradigm is more than relevant to the field of the contemporary artificial intelligence, robotics and cognitive science generally, thus an often shared (and very courageous) idea of the way to create the Golem is doing so by employing a complex system on a certain implementation level, whereas on a "higher" level (perhaps the level of human cognition) the phenomena of a conscious mind of the system "autonomously" *emerge*.

Obviously when designing such a system there should be a concept and a description or causal explanation of the implementation level as well as a theory of how the phenomena of the implementation level actuate and interact with the phenomena of the human cognition level (i.e. the knowledge of the kind: "if one

causes to happen this or this in a certain neural network, the Golem *exhibits* fear"). Of course this "theory" needs not to be explicitly known by the system designers – it can be (and often is) implicitly embodied to the structure of the system. However, the more the causal structure of the system is unknown to its developers, the less likely is the system to be called *artificial* in the sense of the developers' intention to "duplicate" something *natural*.

Eventually if a Golem arises "by a chance" (i.e. coincidentally – as if a lightning hits a mass of jumbled neurons and they suddenly start to perform "conscious" actions; or through the random way of the evolution and "natural" selection), it is hard to treat him as being artificially created. On the other hand, if one theoretically copies a human brain – neuron by neuron – and puts it into an "artificial" body copied "cell by cell" from a human (so as to ensure the being to be *exactly* same as a human and thus experiencing exactly same mental processes – though this is a very courageous premise), it is then again a *human*, not a Golem (and a product of neurobiology and genetics, not artificial intelligence), and the only difference between him and us would be that he is told he is copied (and thus "fake" and artificial) while we are told we are natural.

Further in this text I will present an alternative formal way of describing particular causal aspects of emergent systems. Its results expose a conceptual paradox in the way the emergent systems are understood and make me sceptic to the possibility of the Golem creation.

Along with [1] I will use the term *domain* instead of *level* to suppress the assumption of an underlying hierarchy. Citing from [1]: "Let us call a *causal domain* any segment (or fragment, or component) of reality within the scope of which causal relations appear to be (i.e., are presented to our knowledge as) *manifest* (obvious, apparent), *comprehensible* (intelligible), and *mutually coherent*. Or, more appropriately, they appear to be *more* manifest, comprehensible, and mutually coherent than causal relations *between different* domains are."

Apparently, the definition of a causal domain is vague, but no more than the notion of causality itself. The important feature of such vagueness is that it properly underlies the way the human consciousness understands phenomena of the world it is cast into. However, if we want to formalise such conscious understanding and vagueness, we need a proper formal (e.g. mathematical) apparatus. Vopěnka's Alternative Set Theory (AST) [2], [3] seems to be very suitable for these purposes and its utilisation contributes substantial findings about "artificial consciousness" (and not only about it), as it is shown further in this text.

2 Mathematisation of Vagueness

The idea, philosophical background and phenomenology of the Vopěnka's broad conception resulting – on the field of mathematics – in AST is thoroughly discussed in [3], yet unfortunately still in Czech only. The elaboration [2] of AST is nevertheless available in English too. It is impossible to explain even partially the whole conception on the space limited by this article, thus I will only briefly recapitulate same aspects of it.

The key concept is the notion of *horizon* and *natural infinity* – explicitly specified phenomena, yet bearing themselves the phenomenon of vagueness. In the discourse area of abstract classes and sets a horizon is modelled by a *horizon segment*. To define the horizon segment we need at least the following brief apparatus:

The word *set* refers to what it commonly (i.e. in the classical mathematics) means – precisely (i.e. sharply, non-vaguely) delimited quantity of objects, as a whole meant to be a single object. *Class* stands for a very similar entity, but the quantity of its objects can be delimited also vaguely. A set is a special case of a class. A set is automatically finite, a class can be *naturally infinite*. Vaguely delimited subclass of a sharply delimited set is called a *semiset*. Semisets model the phenomenon of natural infinity (and thus vagueness). Examples of what can be modelled by semisets are: the quantity of sand grains needed to be removed from a hill of sand to be able to say it is not a hill anymore; how many hairs one must remove from a hairy man to make him balding; all my favourite books from the National Library; etc.

Further in the text I will use this notation: $\langle x_1, x_2, \dots x_n \rangle$ is an ordered n-tuple of objects $x_1, \dots x_n$ (in this order); relation R is a class of all 2-tuples $\langle y, x \rangle$ such that objects y and x (in this order) are in a relation R (e.g. yRx); for a relation R symbol $dom(R)$ is a class of all x such that $\langle y, x \rangle \in R$, i.e. domain (in the mathematical sense, not the causal domain); for a relation R symbol $rng(R)$ is a class of all y such that $\langle y, x \rangle \in R$, i.e. range; for classes A and B operator $A|B$ produces a class of $\langle y, x \rangle$ such that $\langle y, x \rangle \in (A|B) \Leftrightarrow \langle y, x \rangle \in A \wedge x \in B$, i.e. restriction. Function F is such a relation where for each $x \in dom(F)$ only one $y \in rng(F)$ is given. This y can be designated $y = F(x)$.

Conjugate function F of a relation R is such a function where $F(x)$ is a set of all y such that $\langle y, x \rangle \in R$. N is the class of all natural numbers (their von Neumann models respectively – i.e. each number is a set comprises of numbers smaller than itself). If $a \in N$, then the class of all natural numbers smaller than a is called a segment of natural numbers, a is its head. If X is a segment, we say a sequence of objects on this segment is stable, provided that for each $a \in X$ a part of this sequence (including a) is delimited sharply. We say f is a solid extension of F defined on X provided that $dom(f) \in N$ (i.e. a sharply delimited segment) and $F = f|X$.

A non-empty class $H \subseteq N$ is a *horizon segment*, iff both these condition are met: 1) in the natural number ordering, H has no last (i.e. biggest) element; 2) if F is a function stable on H, then there exists its solid extension f.

You can see there is no similar structure like this one in the classical mathematics and this is the reason why the classical mathematics is so hardly applicable when discussing phenomenological questions of a consciousness science. The horizon segment as a mathematical structure fairly accurately models the (vague) way a human consciousness meets and treats phenomena of the outer world. And such as it is, we can employ it to model some aspects of how the human mind deals with the phenomenon of emergent systems.

We say a class A is distinctive on a horizon segment H provided that there exists a stable function F on a horizon segment H such that $A = rng(F)$. We say X is a σ-class (respective to H) iff there exists a distinctive class of sets A such that $X = \bigcup A$ (i.e. union of the sets which are elements of A). We say X is a π-class iff there exists a distinctive class of sets A such that $X = \bigcap A$ (i.e. intersection of the sets which are elements of A).

According to [3] we can prove the following *Theorem 1*:

> Be $X \subseteq w$, where w is a set. Be X a σ-class (π-class respectively). Then $(w - X)$ is a π-class (σ-class).

This theorem has a crucial importance, as it will be shown further in the text. There is also *Theorem 2*:

> A class X is a σ-class and a π-class simultaneously iff X is a set (i.e. sharply delimited).

3 Mental States as an Emergent System

We have already set up the notion of causal domain which is of great importance for us, since it represents the way the human consciousness and mind structure the outer world and its functioning. The idea of the artificial intelligence emergentist paradigm is to explain mental phenomena and processes as being caused by complex organisation (and rather simplistic local behaviour) of phenomena and processes of a different causal domain, most often the neuronal one (or evolutionary, molecular, or even sub-atomic). This way the AI emergentist program expects to be able to construct the Golem.

The important aspect of the emergentism is the fact it presupposes the emergent systems can be *in principle* fully described analytically (e.g. by complex non-linear systems), but we just cannot find such description because of the limits of our senses, perceiving and even mental dispositions. In my opinion this relation is not so straightforward – rather I would say our mental and consciousness dispositions are as they are because of the emergence and the emergence exists because of the horizon laid by the limits of our minds. If we go beyond the horizon, the phenomenon which has been emerging on it comes apart and simply ceases to exist (note the closeness of the epistemological and ontological perspectives of this existence).

3.1 Causal Domain Model

Now let us suppose the existence of the causal domain of neural systems (either human brains or artificial neural networks) and the causal domain of mental processes (mental or psychological states and perhaps even the consciousness). For our considerations and purposes it actually does not much matter what *exactly* such domains include – we can understand them in an intuitive way. These domains are a part of far more extensive universe of all conceivable phenomena and

we know both domains are certainly disjunctive, the neural domain includes the phenomena such as the state of various neurotransmitters and synaptic potentials, and the mental domain consists of the phenomena like fear, joy, depression, etc., but concerning both domains we cannot precisely and sharply delimit their borders – we cannot point out a specific phenomenon and say this is the last phenomenon of the particular domain and no other belongs to it.

We can uniquely assign each phenomenon from the universe with a natural number. Albeit this way we empty the essences of the phenomena, we still can explore their behaviour as a quantity. Let a class $U \subseteq N$ be the segment of natural numbers modelling the phenomena universe by the aforementioned unique number assignment. Let a class $H_i \subset U$ be such a segment whose elements are numbers assigned with those phenomena which are placed into the neural causal domain (from this point we will denote this domain by the letter i). For the sake of simplification we can say H_i is a class of the phenomena belonging to the causal domain i.

Apparently H_i is a horizon segment – if we go from the "centre" of the neural causal domain to its edge, we can ask the question whether a certain phenomenon belongs to this domain. If we say it does belong, we know that the next phenomenon ("closest" to the previous one, in the epistemological sense) belongs to it as well. Simply we cannot point out a specific phenomenon and say that this phenomenon is the last one from the neural causal domain. However, we certainly know that we can point out another phenomenon from the universe U which does *not* belong to it (e.g. the air temperature, eruptions on a distant star, someone's sadness, etc.). It is very similar to the situation on the field of scientific disciplines (which can be called "discourse domains"), for example making the (virtual) way from the mechanics to the psychology. Therefore causal (and also discourse) domains apparently bear the phenomenon of natural infinity and thus can be mathematically underlied by a horizon segment.

Let us further suppose the causal functioning within the causal domain i is known, i.e. there is a theory of how the things work inside this domain. It means we know which phenomena (consequences) are caused by a particular phenomenon (a cause). We can model this domain-specific theory by a relation R_i such that: $dom(R_i) = H_i$; $rng(R_i) \in H_i$ and two objects (phenomena) x, y are in the relation R iff x directly causes y (i.e. y is a direct consequence of x). For example, a specific level of the fluoxetin (as a serotonin reuptake inhibitor) presence in the human neural system causes a specific level of the serotonin presence (or the particular sum of the input signals of an artificial neural unit causes a specific signal to be generated). Thus these two phenomena are in the relation R_i.

Let F_i be the conjugate function of the relation R_i – for each $x \in H_i$ the $F_i(x)$ is a set of all consequences of x. The relation R_i is stable on the segment H_i and thus also its conjugate function F_i is stable, hence the class $A_i = rng(F_i)$ is distinctive on the segment H_i. As a result of this, the class $D_i = \bigcup A_i = \bigcup rng(F_i)$ is a σ-class. The class D_i apparently consists of all the phenomena from the domain i which are consequences of some phenomena from the same

domain (i.e. D_i does not contain "initial conditions" – phenomena not having any cause) and I call it a "skeleton of the domain i".

3.2 Inter-domain Causation

Previously I have mentioned we are concerned also in the mental domain (i.e. the domain of mental and psychological processes, states and phenomena, or consciousness, or whatever like this – the name really does not matter). The phenomena of this domain find their models within the universe U too – they can be underlied by a class C_m whereas $H_i \cap C_m = \emptyset$ (i.e. both domains, obviously, have no common elements, otherwise they would not be two *different* domains).

The key step of creating the Golem (as an artificial consciousness) would certainly be finding the causal dependencies *across* those two causal domains (in our concrete virtual experiment; the reality may need much more causal domains to take into account) – i.e. what a certain event or phenomenon in the neural domain causes in the mental domain, or what is a cause of a mental phenomenon in the neural domain.

We take the class of the mental phenomena we are interested in and "join" it with the original neural domain i. This way we obtain a set w such that: $D_i \subset w$ and $w \cap C_m \neq \emptyset$. It means the set w consists of the whole domain i plus it adds a class of relevant phenomena from the mental domain.

Now we move outside the neural domain towards the mental one and from this (natural) perspective we want to see how the phenomena from the class $X_i = (w - D_i)$ are causally dependent on the phenomena from D_i (or H_i respectively).

Analogically to the the function F_i, which describes the causation within the domain i, we want to find a function F_{Xi}, which comprises the inter-domain causal relations between the neural phenomena (from the class D_i) and the mental phenomena (from the class X_i).

And here comes the most important point: according to the *Theorem 1* the class X_i is a π-class and hence $X_i = \bigcap rng(F_{Xi})$ (note the difference – D_i is constructed using the $\bigcup$ operator whilst X_i *must* then be constructed by the $\bigcap$ operator). This means that all the elements of X_i must be included in all the elements of the class $rng(F_{Xi})$ (themselves being sets). Taking the interpretation of the function F_{Xi} into account it means that each mental phenomenon from X_i must be principally included among the consequences of *all* the neural phenomena from H_i!

Informally speaking: if we – as conscious beings, from the scope of the mental domain – ask, which phenomena from the neural level cause particular mental phenomena, the answer is "all the neural phenomena are the cause". This inter-domain causal paradox means the searching for the specific inter-domain causal function is senseless and the aforementioned artificial intelligence emergentist program of the Golem creation cannot be fully successful.

Nevertheless, this paradox points out why unexpected phenomena *emerge* in the emergent systems – they are "unexpected" because we cannot expect them due to lack of their strict causal determination. And according to the *Theorem 2* we can see that an emergent system is only such a system where *vagueness* in

some form is present. The causal paradox allows the emergence by "implicit" adding (since from the ontological point of view some single phenomenon *must* happen at a time point, no matter the epistemological causal "blockades") of "new" phenomena to those expected and "fixed" in the class X_i – the class X_i "grows" in time so as to balance the causal "disequilibrium" posed by the causal paradox. Hence the phenomena of consciousness can emerge this way.

3.3 Causal Paradox Interpretation

As K. R. Popper argues [4], the causality is not an empirical feature of a system, rather it comes to the system along with a theory which describes its functioning. Obviously various theories correspond to various discourse domains which themselves are products of some form of consciousness or a conscious perception respectively. It means the presence of consciousness is necessary to perceive the causality (and not only perceive, but also "make causality exist"). And here we have a hermeneutical circle in its very explicit form: a prior consciousness is essential for the existence of causal domains, but only the causal domains (and their inter-domain causation) can lead to the emergence of the consciousness.

In other words: a conscious mind can be observed emerging on the basis of causal domains, but there must be someone who observes. We can either say that one of the objects (i.e. the conscious mind or the causal domains) was given a priori, perhaps by the God, or we can assert that none of those objects was given a priori, whereas this second option can be explained "evolutionary": the conscious mind and the causal domains have been arising concurrently "step-by-step" towards the immense complexity, as they eventually reached the state we know today.

However, this explanation is also very problematic, since the aforementioned "step-by-step" evolution demands vagueness (eliminating the principle of non-fading induction) and can be again underlaid by a horizon segment and thus leads to the "second-order" emergence, from which the "first-order" emergence emerge. The hermeneutical circle is extended into a hermeneutical "sphere". We can analogically continue and create "higher" emergences together with hermeneutical "hyper-spheres" of any arbitrary order. Moreover, we still must settle on an assumption that at least the phenomenon of the emergence (vagueness respectively) was given a priori. In any case both explanations cross the science boundary and belong to metaphysics.

From this perspective the causal paradox introduced in the previous paragraphs can be interpreted not as a real paradox (in the very sense of this word) but as a fundamental basis on which conscious minds (beings) enact their worlds. In other words: we all know that particular states of our neural system (brain respectively) has something to do with our mental and conscious states and that this relation exhibits regularities (as opposed to complete chaos and randomness). Hence from the ontological perspective we may assume there *is* a "clear" relation between consciousness and neural states.

However, our minds are put (enacted) into such a position from which – as the causal paradox shows – they principally cannot include this relation into their

world understanding. That is for the epistemological point of view and I hope it is obvious how arguable the ontological existence without the epistemological one is. Any attempt to explain the consciousness eventually either explains something different than consciousness, or falls beyond the demarcation line of science not allowing any falsification (i.e. the criterion introduced by Popper [4]).

Moreover, the results shown above can contribute in some way to the discussion about the problem of *free will* and the idea of its blind causal determination: there is no blind causal determination of free will "upwards" from the neural (or whatever else) domain because there is no direct inter-domain causation (in the explanatory sense – i.e. a phenomenon from one level explains a phenomenon from another one) – each state of our neural system causally determines more (and perhaps all) possible conscious states. This can "explain" the Searle's *gap* between a cause and its consequence in the mental domain.

This is nevertheless connected with another problem: if "all" mental phenomena are consequences of "all" neural phenomena, how does it come that at each time point only one mental phenomenon is "chosen", realised and observed, and "who" does then "choose" this particular phenomenon to be realised (i.e. what is the nature of the emergence itself, figuratively explained by "growing" of the class X_i)? This problem can probably be formally explained by the AST modified theory of formal (stochastic) causal systems and by understanding the time continuum in terms of AST. However, this framework still needs to be thoroughly explored and its discussion reaches beyond the extent of this paper.

4 Conclusion

I do not say it is principally impossible to create an artificial being just like a human. However, I intentionally do not use the words "artificial conscious being" and I will explain the reason: the presented causal paradox can "simply" be overcome by reductionist deflating of the mental domain, saying that anything like mental phenomena (love, fear, consciousness, etc.) does not really exist (i.e. they are just epiphenomena) and explaining them only in terms of the neural domain (i.e. "nothing-but-ism"). This reductionist approach, using its "nothing-but" reduction simply removes all the vagueness present in a system and hence removes its emergence and all emergent phenomena (just like those mental ones, but also the free will, the notion of life itself, etc.) – again, see the *Theorem 2* and its discussed consequences.

If there is some vagueness present in the description and understanding of an ant, we can easily speak of this ant in terms of the intentionality (e.g. "the ant wants to get there", "the ant feels the pain and thus runs away", etc.). However, if we reduce the whole ant's behaviour and life to the domain of his neural system functioning, there will soon be no space for any vagueness and suddenly the ant ceases to be a living being and starts to be only a reactive agent, where the notion of "life" has no sense. All living beings are "constructed of" billions of reactive agents and only the presence of vagueness and the inter-domain "paradox causation" makes sense of their lives.

On the basis of what I argued for in this paper we could see that the constitutive element of the consciousness as a phenomenon is a "state of affairs" which I would call a "*fundamental agnosticism*". This fundamental agnosticism is not a mere deliberate epistemological attitude of a person, it is an integral part of the (enacted) consciously perceived world (and manifests itself also in the Gödel's incompleteness theorems).

We can say that conscious minds enact themselves and their worlds by "not seeing everything". One can always extend his sight and thus explain what he saw before as something different. However, what was seen before can never be recalled again. For example the God will never be the same as thousand years ago due to immense changes in the mankind's sight. The question remains whether the way of such sight extending is infinite (theoretically the Gödel's theorems say so) or it eventually collapses. The collapse point can be the theoretical moment when the mental domain becomes empty and the human consciousness is "explained" only in terms of the aforementioned reductionism (although it would be something different than the consciousness what would be "explained"). Yet still the question is whether we can reach such a collapse point: perhaps we cannot reach exactly this point (likewise we cannot reach the "end" of a σ-class) but maybe we can "jump" over it.

Hence if we eventually manage to *create* the Golem, it will be no Golem for us anyway, because there will be no "us" and no consciousness, which is able to judge if that Golem is really *The Conscious Golem*, since we all will be nothing but reactive agents, as the final causal paradox will be removed and the last emergence – giving us the life as we know it – will be lost.

References

1. Havel, I.M., Causal Domains and Emergent Rationality. In: Proceedings of the 23rd International Wittgenstein Symposium "Rationality and Irrationality", Kirchberg, Austria, 2000.
2. Vopěnka, P., Mathematics in the Alternative Set Theory. Teubner, Leipzig, 1979.
3. Vopěnka, P., Meditace o základech vědy. Práh, Praha, 2002.
4. Popper, K. R., The Logic Of Scientific Discovery. Routledge, London and New York, 1995.