MECHANISM AND PERSONAL IDENTITY

Bruno MARCHAL
I.R.I.D.I.A. Université Libre de Bruxelles
50 av. F. Roosevelt. CP194/6. B-1050 Brussels, Belgium

The soul is a number which moves itself.
Xenocrate (see 44)

Abstract : Some thought experiences seem to refute the possibility of subjective experience for machines. By using the recursion theorem of Kleene, I try to invalidate these refutations. A new paradox occurs. I generalize an idea used in the foundation of Quantum Mechanics to suggest a step toward a solution.*

Key words : Machine, Recursion, Duplication, Modality, Quantum Mechanics.

1. INTRODUCTION

I give an intuitive definition of a strong version of Mechanist Philosophy (called simply Mechanism) and I present some paradoxical situations which look like refutation of this philosophy. Then I will be more precise about Machine and sketch a more rigorous "mechanist theory of Identity" based on the Kleene recursion theorem which throws some light on these paradoxes. A new kind of paradox appears then. This paradox bears some relationship to the problem of measurement in Quantum Mechanics, for which there already exists a mechanist solution (see 12, 14, 18, 43). Then I suggest that the present approach generalizes that solution with the consequence that mechanist philosophy would fit both with "Strong A.I." and with quantum mechanical facts.

*The following text presents research results of the Belgian National incentive-program for fundamental research in artificial intelligence initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the author.

2. MECHANISM

The simplest intuitive strong mechanist axiom is the following non metaphorical claim : "my brain is (at some level) a computer". The Mechanist will agree that he is not able to prove that a computer could vehiculate subjective experiences but he feels that he is not able to prove that for a human or even for himself. So his claim will rest on a semi-empirical analysis. The Mechanist does not believe that a neuron (or anything...) is able to collect in one step an infinite amount of information from the neighbourhood so he believes that there is a level at which some digital machines can replace any part of his brain (including biochemical regulatory pathways).
Consequently the Mechanist will admit that there is a level of organization of his brain such that some mechanical substitution of all parts will preserve his identity (see 19, 20, 24, 28, 32).

3. THE PARADOX OF THE POSTPONED DUPLICATION

Mechanism entails that we are duplicable entities. Let us call "translator" a machine which consists of a transmitter-annihilator part and a receiver-builder part. The machine is supposed to be 100% reliable. An (intuitive) mechanist is someone who does not fear to use a translator as a vehicle (although he does fear death). Suppose now that the transmitter-annihilator part is in Brussels and that the receiver-builder part is in Washington. The Mechanist trusts the vehicle and this means that he believes in Brussels that the probability of finding himself in Washington after using that vehicle, is one. What will happen if there are two receiver-builder, one in Washington and

one in Moscow ? It is not difficult, at an intuitive (although Platonist) level to show (by iterating the experience) that mechanism entails that the probability is 1/2. Moreover, a strict subjective indeterminism (inside OR) appears here although from the outside point of view the situation is determinist (outside AND) (see 24). I call that form of indeterminism : mechanical indeterminism.

The original has not privileged status. So if a mechanist uses the translator with a single receiver-builder in Washington and a transmitter *without annihilator* in Brussels, the probability of finding itself in Washington, or staying in Brussels, is still 1/2. Now the instantaneous state description read by the transmitter can be put on a magnetic tape -or a crystal-, which permits postponing the building of the copy. Let us look at the following argument given by Mister X.

Let us call t_1 the instant of capturing my description in the crystal. Suppose $t_0 < t_1 < t_2$, I know, said Mister X, that in the interval $(t_0\ t_1)$ I can postpone in the interval $(t_1\ t_2)$ the decision between using the crystal to build a copy of myself or to destroy that crystal. But this entails that I know during $(t_0\ t_1)$, that I will be the one who will take that decision in $(t_1\ t_2)$, so I know that the probability to remain the original (the one who will take the decision) is 1. (and this reasoning can be used to discourage any use of a translator as a vehicle). The copy will appear to be like me but I will know a posteriori that I will not be him. Note that all the reasoning here can be done in the interval $(t_0\ t_1)$and that makes the argument a priori. Suppose now that at t_2 he decides to build the copy. Mechanism entails, at least, that the copy will say : "Oh, my God! ...I was wrong", and the reconstituted copy will also realize that he will not be able to convince his older self (due to the fact that he knows that he has been convinced by the a posteriori -> a priori argument). Who is right ? Anyway a mechanist (like the reconstituted copy) will have to admit that if there is no backward causation the probability will depend on his self knowledge and in particular on his ability to respect his own decision.

Here is another question : what is the probability in $(t_0\ t_1)$ of staying alive at t_3 with $t_3 > t_2$ if the reconstituted copy is destroyed in the interval $(t_2\ t_3)$? It seems that the probability of being destroyed is 1/2 although the probability of staying alive at t_3 is 1. There are a lot of translator-like paradoxes possible (see 24). The following paradox can be seen as a kind of limit of such thought experiences.

4. THE PARADOX OF THE FILMED TWO-DIMENSIONAL COMPUTER

There exist two-dimensional computer (see 11). Suppose that the low level modules of that bidimensional machine act and react by luminous messages. The machine is embedded in semi-opaque white smoke between two panes of glass. The sensibility of the modules is such that the presence of light suffices to trigger them. So it is possible to activate a module from the outside by light projection. Mechanism entails that it is possible to compile for such a machine an instantaneous state description of a brain of a dreaming person. Mechanism entails that the evolution of the corresponding process on that two-dimensional machine will vehiculate the dreaming experience. Now we can film the evolution of that machine. The question is : does the film vehiculate a subjective experience ? Because there is no more mechanical causation in the film it would seem foolish to expect, even from a mechanist philosopher, an affirmative answer to that question. The trouble is that, relatively to the initial context which here is just the instantaneous state description D (that is why I talk about a dream), the behaviour of the film is always locally equivalent to the behaviour of the two-dimensional computer. It suffices to project the film in real time *and real space* on it, it being reset at the instantaneous description D. We can remove one, two, three … any pieces (modules) of the two-dimensional computer without changing anything. So it seems that mechanism entails that the simple projection of the film generates the experience for that situation is reducible

to the emulation of the two dimensional machines when all parts of the machine have been removed. The trouble is that, for the same reason, we can delete any parts of the movie and, of course, the entire film itself. Does *nothing* emulate dreams ? Because any decision to choose a frontier between what can and what cannot vehiculate the experience, during the two finite removing processes (of the machine's parts and of the movie's parts) seems to be arbitrary, the reasoning looks like a refutation of Mechanism by a reductio ad absurdum. Nevertheless, I argue that such a refutation is not valid (more details are given in 24).

5. IDENTITY

The identity theory sketched here is based on the second recursion theorem of Kleene and its formal (and Platonistic) version known as the diagonalisation lemma (see 1). The embedding of the subject in the object I want to perform depends on the closure of the set of intuitively computable functions (ICF) for the diagonalisation operations. Paradoxes of self-reference are transformed into infinite processes. I will always implicitly use Church's thesis (see 22). The code of a program will be considered here as a necessary body which permits the program to manifest itself relatively to an universal environment. The theorem of Kleene permits us to write programs which are able to output an intuitively computable transformation of their own code. Informally, the idea is the following : it is not difficult to write a program P which, given as input the code ⌜ X ⌝ of a program X, outputs the result of a transformation T applied to the code of a program which compute X(⌜ X ⌝), (first diagonalisation). So P(⌜ X ⌝) outputs T(⌜ X(⌜ X ⌝) ⌝). P has a body ⌜ P ⌝. The result of the application of P on its own code ⌜ P ⌝. : (second diagonalisation) : P(⌜ P ⌝) outputs : T(⌜ P(⌜ P ⌝)⌝). For example, if T is the identical transformation, P(⌜ P ⌝) will be a self-reproducing program. It is a program which builds a copy of itself like a man who uses a translator. The method is constructive. In LISP (for instance) it gives the following self-reproducing *expression* :

```
((LAMBDA (X) (LIST X (LIST (QUOTE QUOTE) X))) (QUOTE (LAMBDA (X) (LIST X (LIST (QUOTE QUOTE) X)))).
```

Some mathematician (see 13 page 227) argue that such expressions are not truly self-referencing because a mathematicien is needed to interpret them. But here we know that a universal LISP program can do the work. The LISP expression above does correctly reproduce itself relatively to a LISP interpreter as a amoeba does correctly replicate itself relatively to natural law. I will insist on Kleene 's theorem by giving a little less informal proof which will give me the opportunity to introduce useful notations. ICF are characterized by the fact that you can define them finitely with finite languages. It is thus possible to enumerate the set of the ICF : ϕ_0, ϕ_1, ϕ_2, ϕ_3, … You can identify the indice i of ϕ_i, with the program which computes ϕ_i. The fundamental properties of such sequences are :

$$\exists u \forall i \forall x \; \phi_u(i,x) = \phi_i(x) \qquad (1)$$

u is a universal program where the code u is able to emulate ϕ_i, and :

$$\exists s \forall i \forall x \forall \mathbf{y} \;\; \phi_i(x,\mathbf{y}) = \phi_{\phi_s(i,x)}(\mathbf{y}). \qquad (2)$$

So parametrization can be automated, with program s. Now Kleene's theorem is (not in his most general form) :

$$\forall t \exists e \forall y \; \phi_e(\mathbf{y}) = \phi_t(e,\mathbf{y})$$

e computes the transformation with code t on itself. **y** denotes a n-uple of parameters. The proof is the same as above : $\lambda x \mathbf{y}.\phi_t(\phi_s(x,x),\mathbf{y})$ (first diagonalisation) is certainly an ICF with variable x, **y**. So there is a r such that :

$$\phi_r(x,\mathbf{y}) = \phi_t(\phi_s(x,x), \mathbf{y}).$$

Using automated parametrization (2) :

$$\phi_r(x,\mathbf{y}) = \phi_{\phi_s(r,x)}(\mathbf{y}).$$

With x = r (second diagonalisation) we get :

$$\phi_t(\phi_s(r,r), \mathbf{y}) = \phi_{\phi_s(r,r)}(\mathbf{y}).$$

So $\phi_s(r,r)$ is our e. (r is playing the role of ⌜ P ⌝ and $\phi_s(r,r)$ the role of :

$$⌜ P(⌜ P ⌝)⌝$$

in the proof described above) (see 22, 29, 33). The proof is constructive and can easily be made uniform (see 5, 25, 29), with obvious s and diag, the second diagonalisation is capture in :

```
(defun k (f)
  (s (diag f) (list (diag f))))
```

Again k applied on identity :

```
(k '(lambda (x) x))
```

gives "the amoeba", a self-reproducing *program* relatively to Lisp (see 23 for details) :

```
(LAMBDA NIL
   (S (QUOTE (LAMBDA (X) (S X
         (LIST X))))
      (LIST (QUOTE (LAMBDA (X) (S X
            (LIST X)))))))
```

I will still need two important sets of results :

5.1. **Self-referential correctness**. If ϕ_e is a theorem prover sufficiently powerful to handle classical elementary propositions of arithmetic (including induction schema), then the above proof of recursion theorem restricted to sentences or formulas is among what ϕ_e can prove (diagonalisation lemma) and it is possible to show that the logic of self-appropriate provable statements obey the modal logic G (see 1, 40, 41) which is a normal system extending K with the axiom :

$$„(„P -> P) -> „P .$$

◊ will abbreviate -„-. T = true, ⊥ = false and „P is an intensional representation of Provable(⌜ P ⌝) and ⌜ P ⌝ is an intensional description of P that ϕ_e is able to handle. Here is the modal version of the second incompleteness theorem in G :

$$◊T -> -„◊T.$$

◊T is -„⊥, that is a consistency statement. Solovay shows that G was complete for self-appropriate provable statements of Peano Arithmetic. Solovay (see 41) proves also that the system G* with all the theorems of G as axioms + the axiom „P –> P, but *without* the necessitation rule is correct concerning G (or extension of G) and is complete for self-appropriate true statement of G (not G*). Concerning sentences and formulas the notion of self-appropriateness corresponds to the notion of self-referential correctness introduced by Smullyan (see 42).

5.2. **Inference inductive machine.** Some Mechanist opponents claim that a computer is an "idiot" because you must program it. To compute ϕ_i you must give him the code i : $\phi_u(i,x) = \phi_i(x)$. We can write i -> ϕ_i, Inductive Inference is the branch of theoretical computer science which works on the inverse process : ϕ_i -> i, i.e. the learning process through examples or phenomena. You give ϕ_i to a machine and the machine tries to find i (or j such that $\phi_i = \phi_j$) (see 3, 7, 16).
More precisely an Inference Inductive Machine (IIM) is a machine which receives successively as inputs, couples <input, output> and which successively outputs programs called hypotheses. The IIM converges if it outputs finally always the same program. An IIM M correctly identifies (or learns or explains) f and we write $f \in EX(M)$ if M converges to a program which computes f (see 7). Note that any ICF is trivially identifiable : ϕ_i is always identified by the constant machine which always outputs (giving any input) i. The interesting concept are the classes of ICF which are identifiable by *one* IIM

M. The collection of such classes is called EX; EX = {L : $\exists$M L $\subset$ EX(M)}. Any set of total ICF which can be generated algorithmatically belongs to EX, but not the whole set of total ICF (see 16). What is interesting is that it is possible to make larger collection of classes of identifiable function by weakening Identification criterion (see 7, 31).

Definition : f$\in$ EX^n(M) if the last hypothesis ϕ_j is such that the number of elements of {x : ϕ_j(x) $\neq$ f(x)} $\leq$ n. We have : EX $\subsetneq$ EX^1 $\subsetneq$ EX^2 $\subsetneq$ …
So by allowing a finite (but bounded) number of errors we get bigger collections of identifiable functions. If the number of errors is still finite but not bounded we get a collection EX* such that $\bigcup_{n\in\omega}$ EX^n $\neq$ EX* although for each n EX^n $\subsetneq$ EX*. If we permit the machine to converge only behaviorally (i.e. : the machine can always change its mind and output an infinity of different programs for all that, eventually, these different programs compute the intended function) we get a collection BC such that EX* $\subsetneq$ BC. And with the same convention BC $\subsetneq$ BC^1 $\subsetneq$ BC^2 $\subsetneq$ …
There is also a collection BC* which includes any BC^i and which is such that $\bigcup_{n\in\omega}$ BC^n $\neq$ BC*. The class of all total ICF does belong to BC*. A beautiful and important result is the following one known as the non -union theorem. I give it for the EX collection but there exist nice generalisations (see 39). There is A, B belonging to EX such that the union of A and B does not belong to EX. This result permits the definition of non trivial identification criteria for collections of machines. (see 9). The recursion theorem plays an important role in the proof of these results.

6. SELF-APPLIED UNIVERSAL IIM

I describe the relation subject/object as a universal machine embedded in a universal machine (think about a n-dimensional cellulor automata). A machine is not able to prove that it is embedded in something much more complex than itself (see 8) so it is a reasonable mechanist assumption (see also 30).
The subject is a program which has a code, or body, or shape. That shape has a "sensitive" surface S "protecting" the code e of an explicit universal inference inductive machine UIIM. The universality of the UIIM means that the machine is able to emulate any hypothesis it synthetizes. e is defined using Kleene's theorem in such a way that the UIIM can repeat, and by inductive inference even anticipate (maybe wrongly), a sufficiently rich set of transformations of the surface S. e is something like <S <U + IIM + e> S>, U is a universal machine emulable by the environment. S can also be seen as a (geometrical) generalization of a READ statement which is the interface between U + IIM + e and the environment. (Those who do not want a mechanical universe must define e = < S < U + IIM + e > S > and ⌜e⌝ = <⌜S⌝ <⌜U⌝ + ⌜IIM⌝ + ⌜e⌝> ⌜S⌝> where " ⌜ ⌝ " are descriptions that e can handle. The reason is that the recursion equation needs to be defined at the soft (or representational level). Such a program learns to emulate what happens to his own code at the surface level (or at the description of the surface level). The output e' of e + change of the surface (correctly reflecting the change of the environment) is e itself including the hypothesis generated abductively by the internal IIM.

$$e \quad e' \quad e''$$
$$e \to e' \to e'' \to \ldots$$

e is able to emulate the hypothesis and also the transformation of his surface. Note that even (especially) if the hypothesis is refutated later, the system e has learned something. When e emulates a change of his surface (waking dream), this must not imply any change of e (although it could), so we must add a flag differentiating at least two levels of emulation (observation itself and the waking dream seen as the emulation of the observation). It is not difficult to introduce another flag which permits the

system to emulate observation with the presence of the waking flag. In that case, the system must be disconnected with the (higher level) surface (sleeping dream).**

7. LOGIC OF INTERNAL IMAGINATION

I will define knowledge of the system from an outside point of view. Basically, knowledge is the collection of what the system is able to emulate. There is a priori nothing *verbal* concerning that knowledge and the fact that the system grows from the learning of the emulation of a surface, entails that this knowledge will have a much more geometrical nature than a logical one. Nevertheless, from the outside, using "–>" for emulate and ⊡ for the (outside name) system, knowledge will be described by the modal system S4. x is an outside view of an hypothesis generated by the IIM and ⊡x represents the partial evaluator when the universal machine of the system fixes ⊡ (parametrization). We have ⊡x→x which means that (⊡x)y = x(y) for any y. (⊡ plays the role of identity or λxy.xy in λ–calculus or combinatory algebra). We also have ⊡x→⊡⊡x which is due to the fact that the system is able to emulate the emulation. This rule would be internalized if we place U, the universal part of ⊡ in the set of hypothesis. But this would entail a lot of trouble (like the knower paradox …) and it would also be contrary to the idea that the system knows only what it learns or experiences. What about ⊡(x–>y)–>(⊡x–>⊡y) : this means that if the system is able to emulate the emulation of y by x (in case x is a more general program than y), then by being able to emulate x, the system is able to emulate y. Modus ponens is evident and the necessitation rule reflects the fact that the system will "know" any events occurring on the surface only if it is able to memorize it, repeat it, emulate it. That system is solipsist. It will be incorrigible concerning local change of the surface and about its ability to emulate hypothesis (independantly of the fact that they will be confirmed or not).***

8. LOGIC OF THE COMMUNICABLE STATEMENTS

Communication acts are necessarily finite if not verbal. I make the hypothesis H that if Platonism is correct Platonist Machines are correct (independently of the fact that they are non referring or even wrong by asserting that they are Platonist (I differ from Putnam here) (see 33)). Note that the Platonism I use is the minimal one which permits me to embed "other (i.e. independent of oneself) mind" in an independent reality. It is the Platonism of a will writer. In that case self-appropriate references which are finitely communicable obey the axiom of G (see 1, 41). But with ⊡P–>⊡⊡P (which reflect internalized induction capability), the system is able, by a (second-order) abduction to infer ⊡-"„"⊥. "„" is the non necessarily

---

** For a precise analysis of the duplication paradoxes, the entire chain "e->e'->e"->e'"->..." must be self-referential and it must admit branching. The following generalisation of the recursion theorem, due to Case (see 6), permit to define such self-referential nets: for all t, it exist e such that:

$$\phi_{\phi_{\cdot\cdot\cdot\phi_e(x_1)}\cdot\cdot\cdot(x_n)}(u) = \phi_t(e,x_1,...,x_n)$$

If the branching is produced by the synthesis of a set of erroneous or approximate hypothesis by inductive inference, then although the nets is presented constructively (outside AND) any "correct branch" cannot be algorithmically determinable (inside OR) (see 25, 26).

*** In 25, I argue that the ultimate (limit) story of a solipsist, in Platonist Mechanist philosophy, is given by S4Grz, i.e. S4 system + the Grzegorczyk formula :

„(„(p->„p)->p)->p

thanks to an elucidation of the relationship between G* and S4Grz provided by Boolos (see 2, 17).

formalized version of „ which exists for self-appropriate machine by hypothesis H. That is not a proof of consistency but a (not ending) experience of consistency. In the limit the machine is able to emulate its own (self)referentially correct) verbal communication. So we will have : ⊡(◊T–>-„◊T); so by (first order abduction) and ⊡(-„◊T) the machine will have ⊡◊T confirming the *feeling* described by ⊡-"„"⊥. Let us read „ P as "I give a convincing communication of", ◊P as "I am able to imagine (or consider, or dream about) P", and by ⊡ "The *system* feels (or knows) that". Then we see that in the limit the system feels or knows that it is able to imagine (or consider) truth but feels or knows that it is not able to give any convincing communication of that fact. If we admit defining consciousness as an internal feeling of consistency (or an equivalent manner : an internal feeling of having the ability to imagine truth) then the approach here explains why solipsism is irrefutable (although Platonistically false). The approach shows also that such machines will eventually have a richer inner set of beliefs than what they will be able to communicate convincingly between themselves. This suggests some epistemic interpretation on the non-union phenomenon. Such machines are also able to refute any attempt to identify them (selves) with a (universal) hypothesis.

9. PHILOSOPHY OF MIND

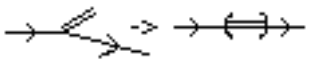
It is not possible, for a sufficiently rich system, to be consistent, self-referentially correct and to obey the reflection principle „p–>p. Now when I propose ◊ for imagination, I am making something dangerous because ◊T is the same as „⊥–>⊥. Of course ⊡◊⊥ is not the same as „◊T but the system obeys ⊡p–>p, so it seems that, at least internally, we will have ◊T. So how could a machine talk about internal logic (feeling) and external logic (convincing communications) without falling into inconsistency by admitting that itself is a machine.

In what logic could the last two paragraphs be formalized ? The solution is : accept "„p–>p", and drop the necessitation rule (or drop the idea that axioms are theorems). Think about an arithmetical theorem prover having a abductive rule to anticipate his own (maybe non stopping) behaviour. Because (like the UIIM) it will prove ◊T–>-„◊T and because „⊥ will never stop, it will in the limit synthesize the metaprogramming rule -„⊥ (experienced) and ◊T (abduction and experience), but it will never be able to infer „(„p–>p). There will never be identification between „ and ⊡ , and the limit result will be G + „p–>p without the necessitation rule, i.e. G*.(both G and G* are finite decidable theory) The arithmetical theorem prover + memory + abductive rules concludes that it is consistent, but because there is a use of the internal experience (abduction, or ⊡ in case of UIIM) it cannot infer that such a conclusion is a communicable one. What it can do is to develop a theory of mind by enriching $S4^{(-)}$ (S4 with a weakening of the necessitation rule) with new axioms. Reinhardt (see 34) proves in $S4^{(-)}$ + a strong mechanist axiom, that there are absolute truths which are not provable. Philosophy of mind is at the G* level, and Philosophy of mind cannot be self-referentially correct. Note the duality between incorrigibility and incommunicability : the first one implies that if you are conscious nobody can "prove" you are wrong, the second one implies that if you are conscious you cannot "prove" it to somebody (see 25).

10. SOME LIGHT ON THE PARADOXES

Naïve Mechanism is correct with respect to the less naïve approach I gave relative to the subject theory I develop. The level at which a translator works is the level which emulates the iterations e–>e'–> … which can all be seen as a "translation"

(or duplication + annihilation). There is no way to give a convincing communication about the very existence of such a level, although it is consistent to admit that there is such a one. Concerning our machines we know from the outside that a level exists (the back-up level!). To choose a level for ourselves depends on empirical consideration. The "probability = 1" argument is at the truth level, that is the G* level (or an $S4^{(-)}$ level) so that in the postponed duplication paradox the copy is right when he said "I was wrong" and the "original" *was* false in his induction . Above all the copy is right when he realizes he will not be able to communicate his error to the original. The experience was just constructed in such a way. During a lapse of time "to be the other", although a true possibility, is actually unbelievable. An UIIM is able to infer that in the iterated duplication without annihilation, the one who always quits the transmitter will in the limit lost his mechanist faith for his experience will be stochastically impossible (see 24).
When copies are destroyed, the argument showing that the probability of being executed is 1/2, although the probability of staying alive in the long run is 1 is also correct. It can be shown, (for exemple in $S4^{(-)}$ + mechanist axioms, where the UIIM is described by a self-referential net as in footnote *** above) that the subjective (solipsistic) experience of the copy can be considered equivalent to a personal forgotten dream :



The filmed bidimensional computer paradox shows, from a UIIM point of view, that any subjective experience is unique and internal. The subjective experience is defined (and redefined) at each state of the process and resumes the entire chain of self-appropriated references, building something like personal history (an internal construct of time) relatively to a universal environment. The paradox arises from a confusion between internal time and external time. The succession of instantaneous state captured by the film does vehicule an experience which has been vehiculated by causal relationship (so there are no new experiences). The paradoxical situation is reduced to a postponed duplication like paradox.

11. THE UNIVERSAL DOVETAILER PARADOX

When a set is such that there is an ICF which generates it, the set is said to be recursively enumerable (RE). The cartesian product of finite number of RE sets is RE. The traditional name given to the algorithm which generates such products is the dovetailer. Having a formal definition of a universal ICF, you can write a program which generates by dovetailing all the finite approximations of all the executions of all the ICF including the presence of any oracles (see 45). I call such a program an universal dovetailer (UD). Oracles play the role of possible environments. The platonist hypothesis (used elsewhere) entails that there is no need for an actual emulation of the UD (for actuality is an inner experience).
Remember that with the theory presented above the existence of a subjective experience corresponds to the existence of chains of appropriate self-reference relatively to an environment : e->e'-> … Let us call such a sequence a sequence of state of mind. For each state of mind the UD is able to process a non denombrable (in the limit) different environment including the most unexpected dreams. Most of them will be inconsistent and will not play any role in the limit (like forgotten dreams). But there is no reason, a priori, that a great number of them can be locally consistent although contradicting the majority of our inductions. The probability aspect of the translator-like paradoxes was easy to work with because the set of possibilities was finite. Unfortunately the Platonist Mechanist solution proposed here entails that we must take all consistent extensions on the actual environment into account. So there is a need to find a measure on the set of sequences of states of mind capable of justifying the

normality of our daily inductive beliefs. Curiously enough such a problem has been partly solved in the context of a realist mechanist attempt to interpret quantum mechanical facts by Hugh Everett and some others (see 12).

12. TURING MACHINES IN A QUANTUM WORLD

In quantum mechanics the state of a system is described by a mathematical object Ψ belonging to a mathematical space H. The evolution of Ψ through time is given by a differential equation S. That evolution is continuous and deterministic.

How to interprete Ψ ? That is a hard and highly debated question. Nevertheless almost everyone agrees on the way to use Ψ. It happens that when we make a measurement of a quantity described itself by a set of some states Ψ_i, Ψ reduces abruptly in a state Ψ_i with a probability computable from Ψ_i and Ψ. That is called the reduction principle. Ψ seems to describe a set of interacting possibilities evolving continuously until we make some measurement, in which case, one possibility occurs. When and how does the reduction occur ? Some have put it at the microscopic level (see 10), others have put it somewhere between the microscopic and the macroscopic level (see 21 for details), still others have put it between mind and the whole physical system including the brain. This last solution has given rise to a lot of rival (almost all dualist) approaches in the philosophy of mind (see 27). I know only one meeting between Mechanists and Quantum Dualists (see 23). There is still the possibility that no reduction occurs at all. The solution of the filmed two-dimensional computer paradox given above is a generalization of that idea. Everett shows by using explicitly the hypothesis that the observer is a machine and that the whole system observer + measuring apparatus + object obeys the differential equation S, that the result of measurement will still reflect the reduction in the memory of the machine. Put in another way, Mechanism entails that the reduction principle can be derived from the continuous evolution S of the whole state of the system in H, provided the measurement is done and interpreted by machines. Does it help to interpret Ψ ? In a sense any universal machine can do that. And any UD does it. But from the inside point of view there is a big price at the ontological (Platonist) level. It has been said that such work is a beautiful theory nobody can believe (see 15) : if the whole environment is described by Ψ, the splitting of possibilities entails the splitting of the environment including the observers, and measurement just tells the observer in which environment he is. Quantum indeterminism is just a particular case of mechanical indeterminism. Propositions about it occur at the G* level and cannot be convincingly communicable. For instance, a common refutation of Everett's work is that it does not explain why we are in this environment and not another. But that remark is, from a mechanist point of view, equivalent to the remark made by the person, who after splitting himself between Washington and Moscow, pretends in Washington (resp Moscow) that Mechanism is incomplete because it does not explain why he finds himself in Washington (resp Moscow). The double-edge nature of the Gödelian argument against Mechanism (see 47) extends itself for the Quantum arguments. For those who knows the Schroedinger's cat paradox, Mechanism + Everett enable us to infer the "subjective experience" of the cat. Roughly speaking, it is the following : "Well, nothing very special, except the presence of more and more physicists with more and more astonished eyes" (resee part **9**). The goal of Everett was to provide an interpretation of Quantum mechanics coherent with cosmology (see 4, 43). Some of these approaches give an equation for the history of the universe in which there is no more explicit reference to time. Here also time is internal and relative. (Information-theoretic extension of Gödel's theorem (see 8) gives hope of finding analogous internal semantics for thermodynamical processes or chaotic dynamics (see 30)). Everett provides not

only a proof that machines do not record the split but also, that if machines make successive measurements they will, in the limit, verify the usual quantum statistics. That work was corrected and refined by Graham (see 12) and independently by Hartle (see 18). It is the equivalent of Graham or Hartle's work which is still needed, concerning the recursion theoretic theory of identity presented here, to solve the UD paradox.

Bibliography

1 BOOLOS G., ***The Unprovability of Consistency, an Essay in Modal Logic***, Cambridge University Press, 1979.

2 BOOLOS G., *Provability, Truth, and Modal Logic*, Journal of Philosophical Logic, 9, pp. 1-7, 1980.

3 BLUM L. & BLUM M., *Toward a Mathematical Theory of Inductive Inference*, Information and Control 28, pp. 125-155, 1975.

4 BROUT R. et ENGLERT F., *Cosmologie quantique*. in **Aux confins de l'univers**, Schneider, pp.269-289, 1975.

5 CASE J., *A Note on Degrees of Self-Describing Turing Machines*, Journal of the A.C.M., vol. 18, n° 3, 1971, pp 329-338.

6 CASE J., *Periodicity in Generations of Automata*, Mathematical Systems Theory. Vol. 8, n° 1. Springer Verlag, NY, pp. 15-32, 1974.

7 CASE J. & SMITH C., *Comparison of Identification Criteria for Machine Inductive Inference*, Theoretical Computer Science 25, pp. 193-220, 1983.

8 CHAITIN G.H., ***Algorithmic Information Theory***. Cambridge Tracts in Theoretical Computer Science 1, Cambridge University Press, 1987.

9 DALEY R.P., *Inductive Inference Hierarchies* : *Probabilistic vs Pluralistic Strategies*, Mathematical Methods of Specification and Synthesis of Software Systems, Springer Verlag, Lecture Notes in Computer Science 215, 73-82, 1986.

10 DE BROGLIE L., ***La théorie de la mesure en mécanique ondulatoire***. Paris, Gauthier Villars, 1957.

11 DEWDNEY A.K., ***The Armchair Universe***, W.H. Freeman and Company, N.Y., 1988.

12 DE WITT B.S. & GRAHAM N., ***The Many-Worlds Interpretation of Quantum Mechanics***, Princeton Series in Physics, Princeton Univ. Press, 1973.

13 ENDERTON H.B., ***A Mathematical Intoduction to Logic***, Academic Press, 1972.

14 EVERETT H., *The theory of the Universal Wave Function*, In De Witt B.S. & Graham N.pp. 3-140, 1973.(see 12).

15 GARDNER M., ***Time Travel & Other Mathematical Bewilderments***, W.H. Freeman and Company, N.Y., 1987.

16 GOLD E.M., *Language Identification in the Limit*, Information & Control 10, pp. 447-474, 1967.

17 GRZEGORCZYK A., *Some relational systems and the associated topological spaces*, Fundamenta Mathematicae, LX pp. 223-231, 1967.

18 HARTLE J.B., *Quantum Mechanics of Individual Systems*, American Journal of Physics, vol. 36, n° 8, pp. 704-712, 1968.

19 HOFSTADTER D., DENNETT D.C., (eds) ***The Mind's I***, Basic Books, inc. Pub., N.Y., 1981.

20 HUME D., ***Treatise of Human Nature***, 1739.

21 JAMMER M., ***The Philosophy of Quantum Mechanics***, J. Wiley & Sons, N.Y., 1974.

22 KLEENE S.C., ***Introduction to Metamathematics***, P. Van Nortrand Comp. Inc., 1952.

23 LETOVSKY S., *Ecclesiastes : A Report from the Battlefields of the Mind-Body Problem*, A.I. Magazine, vol. 8, n° 3, pp. 63-69, 1987.

24 MARCHAL B., *Informatique théorique et philosophie de l'esprit*, Acte du 3ème colloque international Cognition et Connaissance, pp. 193-227, Toulouse, 1988.

25 MARCHAL B., *Amoeba, Planaria and ...Dreaming Machines*, submitted to publication, 199?.

26 MARCHAL B., *Des Fondements Théoriques pour l'Intelligence Artificielle et la Philosophie de l'Esprit*, Revue Internationale de Philosophie, 1, n° 172, pp 104-117, 1990.

27 MARGENEAU H., ***The Miracle of Existence***, Shambhala, New Science Library, Boston & London, 1984.

28 MINSKI M., ***The Society of Mind***, Simon and Schuster, N.Y., 1985.

29 MYHILL J., *Abstract Theory of Self-Reproduction*, in Views on General Systems Theory, M.D. Mesarovic, ed. Wiley NY, 1964, pp 106-118.
30 Mc CAULEY J.L., *Chaotic Dynamical Systems As Machines Computational Systems - Natural and Artificial*, Springer, Berlin, 1987.
31 OSHERSON D.N., STOB M., WEINSTEIN S., ***Systems that learn***, The M.I.T. Press, Cambridge, 1986.
32 PERRY J., (ed), ***Personal Identity***, University of California Press, Berkeley, 1975.
33 PUTMAN H., ***Reason, Truth and History***, Cambridge University Press, 1981.
34 REINHARDT W.N., *Epistemic Theories and the Interpretation of Gödel's Incompleteness Theorems*, Journal of Philosophical Logic 15, pp. 427-474, 1986.
35 ROGERS H., ***Theory of Recursive Functions and Effective Computability***, McGraw-Hill, 1967.
36 ROSLER O.E., *Endophysics*, In Casti & Karlqvist (ed), Real Brains, Artificial Minds, North Holland, 1987.
37 RUCKER R., ***Infinity & the Mind***, The Hervester press Ltd, 1982.
38 SHAPIRO S., ***Intensional Mathematics***, North Holland, 1987.
39 SMITH C.H., *The Power of Pluralism for Automatic Program Synthesis*, Journal of the Association for Computing Machinery, vol. 29, n° 4, pp. 1144-1165, 1982.
40 SMORYNSKI C., ***Self-Reference and Modal Logic***, Springer Verlag, 1985.
41 SOLOVAY R., *Provability Interpretations of Modal Logic*, Israel Journal of Mathematics 25, 1976, pp. 287-304.
42 SMULLYAN R.M., *Modality and Self-Reference*, In Shapiro, 1985, pp. 191-211, (see 38).
43 TIPLER F.J., *Interpreting the Wave Function of the Universe*, Physics Report (review of Physics Letters),137, n° 4, pp. 231-275, 1986.
44 TROUILLARD J., ***L'un et l'âme selon Proclos***, "Les belles lettres", Paris, 1972.
45 TURING A.M., *Systems of Logic Based on Ordinals*, Proceedings of the London Mathematical Society, ser. 2, vol. 45, 1939.
46 WANG H., ***From Mathematics to Philosophy***, Routledge & Kegan Paul, London, 1974.
47 WEBB J.C., ***Mechanism, Mentalism & Metamathematics, an Essay on Finitism***, D. Reidel Pub. Company, 1980.