Algorithmic Social Sciences Research Unit

ASSRU

Department of Economics
University of Trento
Via Inama 5
381 22 Trento Italy

Discussion Paper Series

16 – 2011/II

NEWCOMB'S PARADOX: A SUBVERSIVE INTERPRETATION♣

K. Vela Velupillai

October 2011

> "Now for an unusual suggestion. It might be a good idea for the reader to stop reading this paper at the end of this section (but do, please, return and finish it), mull over the problem for a while (several hours, days) and then return. It is not that I claim to solve the problem, and do not want you to miss the joy of puzzling over an unsolved problem[1]. It is that I want you to understand my thrashing about."
> Nozick (1970), p. 117

♣Without intending to help me out with a conclusion that is reasonable from the point of view of the poor mortal Player, pitted against a Demon, my critical friend Alfredo Pastor – a personification of *reasonable man* – provided a reaction which, in fact, seemed obvious. He is, of course, not responsible for my interpretation of his reasonable reaction to an earlier, somewhat deliberately *incomplete*, version of this paper.

[1] The way I have reformulated the problem, it can be made into an 'unsolvable problem', in the precise sense made famous in Turing (1954).

Abstract

A re-interpretation of the asymmetric roles assigned to the two agents in the genesis of Newcomb's Paradox is suggested. The re-interpretation assigns a more active role for the 'rational' agent and a possible *Turing Machine* interpretation for the *behaviour* of the demon (alias 'being from another planet, with an advanced technology and science,..,etc.'). These modifications, while introducing new conundrums to an already diabolical interaction, do allow the 'rational' agent, as a *computably behavioural agent*, to make a clear decision, if any decision is possible at all. This latter caveat is necessary because in the Turing Machine formulation, the computably behavioural agent might have to face *algorithmic undecidabilities*.

§1. A Brief Preamble

William Newcomb devised the Newcomb Paradox in 1960, but it was left to Robert Nozick (1970) to revive it in the context of choice theory and make it one of the staples of the conundrums facing the decision theorist, almost in the class of the more famous Allais, Ellsberg and other similar paradoxes.

To the best of my knowledge, most – if not all – of the proposed 'solutions' have left the 'Being' – who will be referred to as *Newcomb's Demon* [3] in this note – unscathed[4]. All of the difficulties of extricating oneself from the Demon's wiles and flawless predictive powers are placed squarely in the court of the 'player' – clearly an ordinary 'mortal'.

I want to suggest a mode of viewing the paradox which could, instead, leave Newcomb's Demon in a dilemma of a *Gödelian* type. This enables the player, at last, to exercise a plain and simple kind of behaviour and walk off with $1000!

§2. The Paradox

Suppose you are playing a 'game' with Newcomb's Demon, who has powerful predictive powers. More precisely (Nozick, op.cit,p. 114):

> "You know that this being has *often* correctly predicted your choices in the past (and has never, *so far as you know*, made an incorrect prediction about your choices), and furthermore you know that this being has *often* correctly predicted the choices of other people, *many* of whom are similar to you, in the particular situation [of this paradox]. One might tell a longer story, but all this leads you to believe that *almost certainly* this being's prediction about your choice in the situation to be discussed will be correct."

[3] In the noble tradition of *Maxwell's Demon*, *Laplace's Demon* and *Hilbert's Demon* in post-Newtonian scientific thought experiments – and not unrelated to the *Walrasian Auctioneer* in economics, who was so named in an unfortunately misunderstood analogy with Maxwell's Demon by Axel Leijonhufvud.

[4] The reference list, below, collects a few of the more famous contributions towards a discussion, explanation and possible resolutions of the paradox. It is not meant to be exhaustive.

I am afraid the statement of the problem is somewhat imprecise – perhaps that was Nozick's intention, from the outset. '*Often*', '*so far as you know*', '*many*' and '*almost certain*'[5] remain open to interpretation in the particular formal system in which one may choose to resolve the paradox. For example, it is possible to interpret the first criterion – '*often* correctly predicted your choices in the past (and has never, *so far as you know*, made an incorrect prediction about your choices)' – as the behaviour of a Demon who only predicts when he *knows* for certain that it will be correct. This will call into question the meaning of 'knows' – but let that pass.

This particular behaviour by the Demon is entirely interpretable – without any probabilistic underpinning – as that by a Turing Machine (or any of its equivalents by the Church-Turing Thesis) facing recursively enumerable sets that are not recursive, in conjunction with the theorem of the *Halting Problem for Turing Machines*.

Before expanding on such an interpretation, let me summarise the paradox. There are two boxes in front of you: one is opaque; the other is transparent and you can plainly see an authentic $1000 note placed in it.

However, you know that *Newcomb's Demon* has placed $1,000,000 or nothing in the opaque box – depending on what the Demon has predicted YOU will do, when your turn comes to choose one of two courses of actions:

1. You can either take the opaque box
2. Or you can take both boxes.

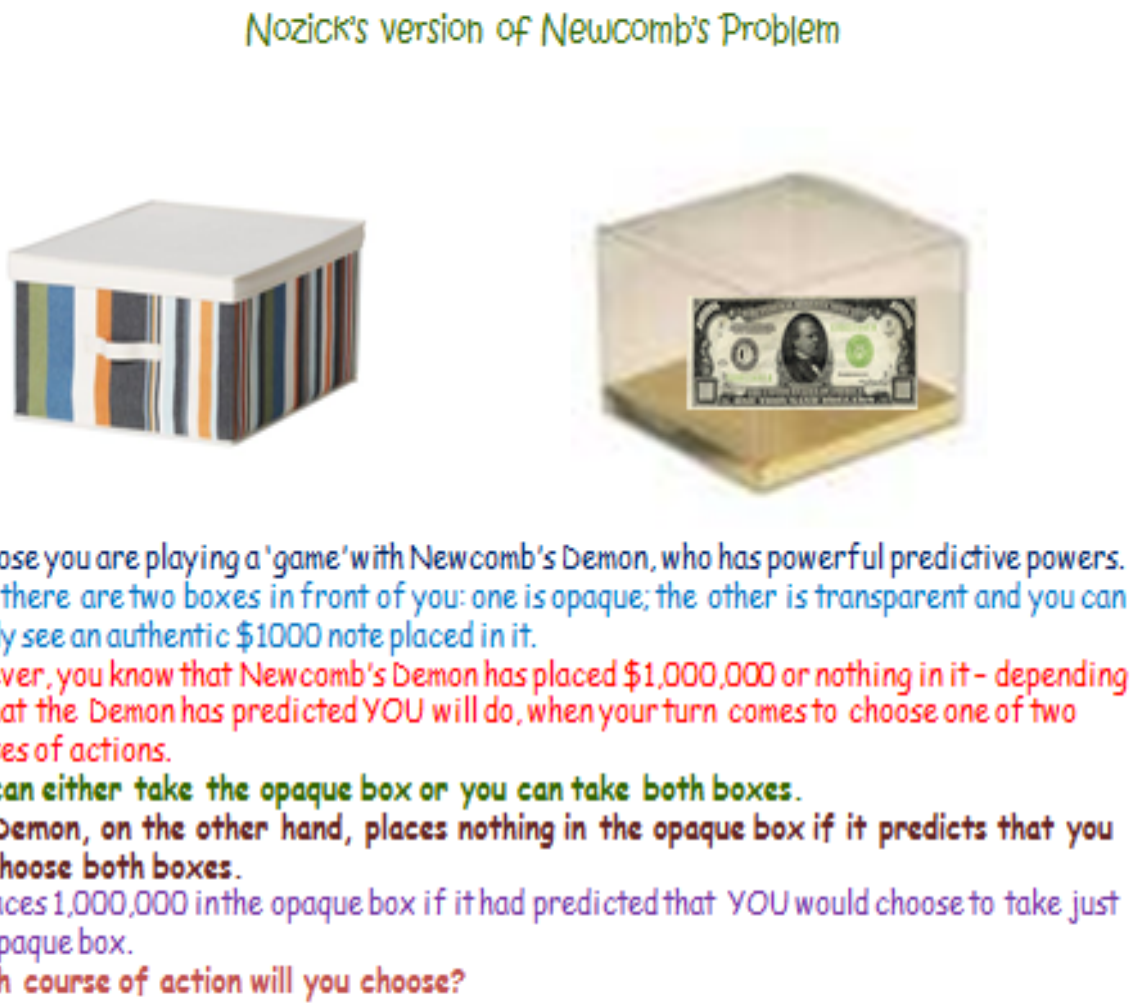
The Demon, too, seems to act in one of two possible ways:

I. It places nothing in the opaque box if it predicts you will choose both boxes.
II. It places 1,000,000 in the opaque box if it predicts YOU would choose to take just the opaque box.

[5] That this notion is, possibly, probabilistically grounded is evident from the discussion on p. 115 and p.131 and the related footnote 16 on p. 145. But it is not clear from the context what kind of probability is being used; indeed, it is not clear to me that Nozick does not freely switch from 'objective' to 'subjective' notions of probability quite arbitrarily.

Which course of action will you choose?

Here is a somewhat exaggerated graphic display of the problem:



A possible payoff matrix for a 'rational choice theoretic' interpretation of YOU vs. the Demon is as follows:

	Demon predicts $\mathbf{P_1}$	Demon Predicts $\mathbf{P_2}$
Action $\mathbf{A_1}$ by YOU	**$1,000,000**	**$0**
Action $\mathbf{A_2}$ by YOU	**$1,001,000**	**$1,000**

$\mathbf{P}_1$: The Demon predicts you will take only the opaque box.

$\mathbf{P_2}$: The Demon predicts you will take both boxes.

$\mathbf{A_1}$: YOU take only the opaque box.

$\mathbf{A_2}$: YOU take both boxes.

Nozick devised this variant of the *Newcomb Demon* problem to highlight the conflict between relying on the principle of maximizing *subjective expected utility* (SEU), and the *dominance principle*. Illuminating discussions and possible resolutions of the paradox can be found in Bar-Hillel & Margalit (1972), Gardner (1973), Horgan (1981) and Levi (1982).

§3. A Brief Recursion Theoretic Note

Two important implicit assumptions in any formal consideration of a resolution of Newcomb's Paradox need to be made explicit before any reasonable solution can even be considered meaningful: *honesty* to one's own thought processes, which I shall refer to as the *Self-Honesty-Postulate* (*SHP*), and *time*.

Suppose Newcomb's Demon, or YOU – or both – are actually, Turing Machines (perhaps in disguise). Essentially this should be analysed as an *Alternating Arithmetic Game* (cf. Velupillai, 2000, chapter 7). Such games may be able to determine the winner in such two person complete information games, but will not provide a method – an algorithm – for the determined winner to implement to 'win'.

But the alternation is in the sphere of *thought* processes and would require the application of recursion theoretic fixed points to resolve a problem of infinite regress intrinsic in such contexts[6].

What does it mean for a Turing Machine to be honest to its own thought processes – or even for YOU to be honest to YOUR own thought processes, supposing YOU are not a Turing

[6] As in the description of the choice YOU make between two actions, Nozick (pp. 114-5):

> "[Y]ou know [that you] have choice between two actions. Furthermore, .. you know this, the being knows that you know this, and so on."

Machine? The simple answer to the former alternative will be analogous to the kind of postulates underpinning a computing machine that has to process *a true sentence that it cannot print, given that the sentences it prints are all true* (cf. Smullyan, 1972, pp. 2-4).

§4. The Demon's Dilemma

Suppose we look at the paradox from the point of view of the Demon, rather than as one in which the rational agent must resolve a choice theoretic dilemma.

The demon is said to examine YOU, for its predictive purposes. What does it *mean* to state that the demon examines YOU, for *its* predictive purposes? One possible thought experiment of the demon's 'examination process' is to suppose that it lists all YOUR feasible thought processes, stated formally as – say – well-formed-formulas(WFF) in a well-defined alphabet. These can be number-theoretically encoded effectively, for example using the usual technique of Gödel numbering.

Let us continue to suppose the demon is a computing machine[7], essentially a Turing Machine, that considers only the feasible thought processes, stated as WFFs. The demon's 'infallible predictive record' could be interpreted to mean that it will only act on the basis of processing one of the possible *Gödel numbers*, for its predictive purposes.

Can YOU induce the demon, through YOUR thought processes, into paralysis? One possible way YOU can do this is to include the following thought in your set of WFFs:

> ***I predict** that the demon will place $1,000,000 in the opaque box if, and only if, it **thinks, you think,** it will not place it in it.*

Call this ***SHP*** (for the *Self-Honesty-Postulate*).

Now suppose the list of your thought processes form a recursively enumerable set that is not recursive. What does the demon do, under the constraint of having to respect SHP and having to check, as a Turing Machine, a list forming a recursively enumerable set that is not recursive ? Now, add the further supposition that the Demon, in its Turing Machine

[7] If, in the theory of evolution, replication is occasionally 'explained' by invoking an analogy with a 'slightly faulty copying machine', I see no undue stretch of the imagination being called forth for an analogy that equates Newcomb's Demon with a Turing Machine.

incarnation, has a time constraint imposed upon its decision process. Can the demon, as a Turing Machine, avoid getting into a non-halting state?

More sophisticated versions of such paralysis-inducing thoughts by YOU can easily be constructed, all of which will challenge *an implicit hypothesis* in the characterisation of an 'infallible' demon: that *the demon is honest to its thought* – i.e., *to itself* – and a time constraint.

In the original version of the paradox, as re-interpreted by Nozick, it is assumed that the demon's action *precedes* the choice by the rational agent[8]; but also that the demon's actions are informed *by its own thought processes*. These latter, in turn, are induced by an 'inspection' of YOUR thought processes.

Therefore, now YOU have the possibility of manipulating YOUR thought processes with the intention of paralysing the demon's action.

In the face of the Demon's possible paralysis, YOU simply take both boxes and walk away with a smug smile on your face!

§4. Brief Concluding Discussion

I conjecture that the most fruitful way to resolve Newcomb's paradox is to remove the asymmetry between the rational agent and the demon and endow the latter, too, with self-inflicted, paradox-inducing, capabilities of (algorithmic) rationality. This will, of course, lead to a new kind of indeterminacy, which is best understood in terms of machine models of thought and action.

In fact, I go further and claim that the Newcomb Paradox can be interpreted most illuminatingly also as a version of the '*Gödelian Puzzle*' in Smullyan (1992), pp. 2-3. There

[8] As described by Nozick (ibid, p. 115; italics added):

> "The situation is as follows: *First*, the being makes its predictions. *Then* it puts the [$1 000 000] in the second box, or does not, depending upon what it has predicted. *Then* you make your choice."

are many variants of such puzzles, all of them resulting in one or another kind of undecidability or unsolvability.

Again, as suggested above, it may be most useful to view Newcomb's paradox dynamically, as an alternating Turing machine game, where the question of the resolution of the paradox is posed as one of *effective playability*. Alternatively, my conjecture is that the paradox could be transformed into a *Diophantine decision problem*.

I cannot but imagine that the resolution, in the former case, will be, appealing to some version of the *undecidability of the Halting problem for Turing machines*, that the game has no effective procedural solution. In the latter case, using the negative answer to Hilbert's Tenth Problem, an alternative form of an algorithmic undecidability could be derived.

In any case, the problem is one that is inherently dynamic and the time element has never been considered in any of the several resolutions of the paradox that have been proposed till now. Surely, the demon should be freed from the possibility of entering a state of paralysis by being given a *time constraint* for its action? YOU can, however, still induce a Turing Machine demon to enter into a non-halting state, which will be the Machine equivalent of entering a state of paralysis with respect to action – and YOU can know that you have induced it into such a state.

References

Bar-Hillel, Maya & Avishai Margalit (1972), *Newcomb's Paradox Revisited*, **The British Journal for the Philosophy of Science**, Vol. 23, No. 4, November, pp. 295-304.

Gardner, Martin (1973), *Free will revisited, with a mind-bending paradox by William Newcomb - Mathematical Games*, **Scientific American**, Vol. 229, No. 1, July.

Horgan, Terence (1981), *Counterfactuals and Newcomb's Problem*, **Journal of Philosophy**, Vol. 78, No. 6, June, pp. 331-356.

Levi, Isacc (1982), *A Note on Newcombmania*, **Journal of Philosophy**, Vol. 79, No. 6, June, pp. 337-342.

Nozick, Robert (1970), *Newcomb's Problem and Two Principles of Choice*, in: **Essays in Honour of Carl G. Hempel**, edited by Nicholas Rescher, pp. 114-146, Synthese Library, Dordrecht, Holland.

Smullyan, Raymond. M (1992), **Gödel's Incompleteness Theorems**, Oxford University Press, Oxford.

Turing, A. M (1954), *Solvable and Unsolvable Problems*, pp. 7-23, in: **Science News**, # 31, edited by A. W. Haslett, Penguin Books, London.

Velupillai, Kumaraswamy (2000), Computable Economics, Oxford University Press, Oxford.