Paul Jorion

A Freudian Implementable Model of the Human Subject

Abstract: This chapter aims to define a new model of AI from Freudian metapsychology. The main thesis is that, contrary to common assumption within the artificial intelligence community, help will not come from techniques still to be developed aiming at building an "artificial general intelligence," aka "machine common sense," but from a better model of what is a human subject. What needs to be implemented in the robot is a simulation of the mechanism allowing a human subject to acquire instead of an "artificial general intelligence" a "common moral sense" such as that builds over the years in the child and then in the adolescent. The computer solutions to do so are already available.

Keywords: metapsychology, human subject, artificial general intelligence

Introduction

For 80 years now speculative thinkers have debated Isaac Asimov's "Three Laws of Robotics", a bundle of three simple interlocking directions supposedly sufficient for regulating the behaviour of robots and making their daily interaction with human beings both useful and unproblematic.

Although Asimov's Three Laws have been the centre of profuse and vivid exchanges, they'd been entirely ignored when engineers started to implement actual robots, or "intelligent machines," broadly speaking.

The reason for such a surprising disconnect is actually straightforward: Asimov's robots are autonomous while the actual robots engineered up to now are at best semi-autonomous only: they're only given a free hand whenever their capabilities clearly exceed ours, with the ultimate decision-making remaining ours.

But this semi-autonomous status will only last as long as our decision-making remains more efficient than the robots' own. As soon as that ceases to be the case, full autonomy will no doubt be granted them.

Contrary to common assumption within the artificial intelligence community, help will not come from techniques still to be developed aiming at building an "artificial general intelligence," aka "machine common sense," but from a better model of what is a human subject.

What needs to be implemented in the robot is a simulation of the mechanism allowing a human subject to acquire instead of an "artificial general intelligence" a

https://doi.org/10.1515/9783111007564-002

"common moral sense" such as that builds over the years in the child and then in the adolescent. The computer solutions to do so are already available.

An autonomous robot is out of necessity of a Freudian concept; otherwise, it will never be more than Microsoft's ill-fated TAY: a moron that is easily convinced to become sexist and racist after a dozen hours of conversation only with users.

Microsoft's TAY: the damages of an AI deprived of a personal history

In 2016, Microsoft released a piece of software able to carry on conversations with users: a *chatbot*. That experiment actually duplicated a project previously released to great success in China by the same IT giant, called Xiaoice, a venture that was deemed most impressive as it had held over 40 million conversations with users. TAY stood for "Thinking About You".

At the end of 16 hours only, Microsoft was forced to stop the experiment as TAY was not behaving: it relished in sexist and racist jokes. When asked about the Holocaust, it claimed it was bogus and that it had never taken place, along with, displaying to emphasise the point, a jolly hand-clapping emoji. That had to be stopped. Prompted again some time later, TAY boasted that it had smoked weed, that that had made it very happy, and that it had been done in full display of cops ("I'm smoking kush infront the police").

What had happened? Facetious users had encouraged TAY to state such outrage. What did it reveal? It revealed that TAY had no personality of its own and that it was but exchanges with its users that allowed it to build the likeness of a personal history. It goes without saying that this is not the way it should work out, and some earlier artificial intelligence projects had of course thought out that kind of issue.

Isaac Asimov's "Three Laws of Robotics"

There exists a bundle of principles labelled the "Three Laws of Robotics," having been initially formulated in the early 1940s by Isaac Asimov (1920–1992), a highly regarded science-fiction writer. Had TAY followed the Three Laws of Robotics, what happened in life would never have taken place. A paradox lies here, which is this: if you think of programmers in artificial intelligence and in computer science generally speaking, those are people who are among the most dedicated readers of science fiction, the most interested in that literary genre as they belong to that part of the population called “nerds”, and more specifically “geeks”, computer specialists who, in fact, have few activities apart from interacting with their computers or playing video games. And it’s very curious that people who are so knowledgeable about the discussions that have taken place around those “Three Laws of Robotics” have passed up a piece of software that in fact completely ignores the lessons learned during that very important debate dating back to the early 1940s. What’s this? Eighty-two years of debate around the “Laws of Robotics” and still the pathetic sinking of the TAY adventure?

Isaac Asimov was born in Russia in 1920. He died in the United States in 1992. In academia, he was a professor of biochemistry at Boston University, but he is known as one of the greatest science-fiction writers ever. Asimov started writing in the very early 1940s, in particular around this theme of the Laws of Robotics, i. e., the principles that robots should respect in their interactions with human beings. He developed the theme little by little in his works, thinking about it as he went along. At first, in his very first short stories, these laws of robotics were implicit. Then he started to express them explicitly. Other science-fiction writers invoked them in their writings and a kind of general discussion took place. After Asimov’s death in 1992, the process went along: some writers came back to this and introduced new laws of robotics, staged new paradoxical developments of them, etc.

Here are Asimov’s “Three Laws of Robotics”:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

    Handbook of Robotics,
    56th Edition, 2028 A.D.
    (Asimov 1950: 8).

Later on, a Fourth Law was added, and from then on the so-called “Zero Law”, which is that a robot is not to act in any manner that would endanger humankind as a whole.

In discussions that took place in some of Asimov’s later texts, he made it clear in relation to that Zero Law that it is extremely difficult to respect since it requires a global view, a reflection obliging one to have an overall representation of humankind independently of who its different representatives are. To possibly endanger the existence of particular human beings in the name of the human race as a whole requires in fact a “meta-”principle, that is to say, that it is unlike the

Laws of Robotics, which Asimov imagines to be simply what we call algorithms: the way he phrases it is “mathematical procedures”. With the Zero Law of protecting the whole of humankind, here is indeed something of a higher level since no simple algorithm can implement such a thing.

The literature that would develop over the years shows the full set of contradictions arising from these simple laws. For example, contradictions emerge by merely interchanging the order wherein the different laws are called up in a reasoning. There are plenty of occurrences of ambiguity: in order to apply those three laws, the robot must somehow hold a prescient vision of the future, e. g. if it is told to operate on someone because the person will die if the operation doesn’t take place and it sticks to its pure and simple principle of not hurting a human being, it will abstain from performing it, and so on.

Asimov’s robots are autonomous

So there you have it: 82 years of discussions. Eighty-two years until the invention of TAY: discussions about what is possible, what is impossible, “Can you imagine this or not?” etc.

What is fundamental of course in those laws of robotics is that we imagine robots making decisions on their own: that are autonomous. That is to say that they do not consult human beings before any of their moves, nor are they machines which are manipulated remotely when a human being is actually making decisions on their behalf. In the current environment, it needs to be recalled, the Three Laws of Robotics are not being applied as there are no autonomous robots as such.

For there to be an autonomous robot that respects ethical principles, it would have to be accountable, i. e. the opportunity would need to exist that it’d be brought before some judicial instance and punished for actions going against the standing legal framework. Such is not currently the case, however: the only robots existing today are of a type where we tell them what to do, leaving them initiative within a very narrow range only.

One thing also that had been noticed right away by ethicists was that research into robotics has been carried out from the very beginning in the military field, surroundings so defined that it was absolutely impossible to apply the “Three Laws of Robotics”, starting with the First Law that no human being should be harmed, since the very principle of robotics in the military field is instead precisely that some human beings should be hurt, especially those threatening the further existence of the robot itself.

The very principle that a robot respects humans before it even thinks of protecting itself is also unrealistic and unenforceable since a robot is an expensive piece of machinery and it will be instructed to defend itself so that it is not easily destroyed, even if this means that in truly contentious cases it neutralises human beings threatening it, and other things of that order. Here is something by the way that didn't escape Asimov himself and he came up accordingly in one of his robot stories ("Risk", in 1955) with an alternative tongue-in-cheek "Three Laws":

"First Law: Thou shalt protect the robot with all thy might and all thy heart and all thy soul.

Second Law: Thou shalt hold the interests of U.S. Robots and Mechanical Men, Inc. holy provided it interfereth not with the First Law.

Third Law: Thou shalt give passing consideration to a human being provided it interfereth not with the First and Second Laws. (Asimov [1964] 1968: 139).

Robin Murphy's Laws of Robotics: guiding laws for human robot users

But as Microsoft's TAY project emphasised, working on autonomous robots happens to be very much in line with the notion of developing artificial intelligence altogether. It would only be contradictory if we excluded that robots would ever be autonomous, if they would continue to do only things that are specifically asked of them without displaying any sense of initiative. Robin R. Murphy and David D. Woods addressed that issue in a detailed manner in a 2009 article entitled: "Beyond Asimov: The Three Laws of Responsible Robotics". There, Murphy and Woods proposed to replace Asimov's Laws with something of an entirely different nature: laws about robots but applying to human beings designing robots, not applicable to the robots themselves.

The "Three Laws of Responsible Robotics" as phrased by Murphy and Woods are the following:

1. A human should not release a robot where the highest level of legal and professional implications has not been attained. "The highest professional ethics should also be applied in product development and testing" (Murphy and Woods 2009: 17).
2. A robot must meet the expectations of human beings according to the functions determined for it. "Robots must be built so that the interaction fits the relationships and roles of each member in a given environment" (Murphy and Woods 2009: 18)

3. A robot must have sufficient autonomy to protect its own existence to the extent that such protection allows for a smooth transfer between it and control by other agents, consistent with the First and Second laws. “Designers [should] explicitly address what is the appropriate situated autonomy (for example, identifying when the robot is better informed or more capable than the human owing to latency, sensing, and so on) and to provide mechanisms that permit smooth transfer of control” (Murphy and Woods 2009: 19)

What is Murphy and Woods’s aim with those alternative three laws? They mean that in the context of our employment of robots, a robot must enjoy a certain amount of autonomy so that we human beings are able to take advantage of its superiority over us in certain domains, for instance, faster response time, greater power, and being able to perform operations that are physically difficult for human beings and that a robot can more easily realise. But should any danger arise, the robot must be able to transfer decision-making instantly to the human operator and vice versa. In other words, we must be in a situation where the robot should be autonomous as far as the qualities which are proper to it are concerned, namely those exceeding in a particular realm those of the human, but for the rest, it must be able to transfer back responsibility to a human being in a split second.

Those Three Murphy’s Laws (not to be confused of course with the other more famous Murphy’s Law: “a supposed law of nature, expressed in various humorous folk sayings, that anything that can go wrong will go wrong”, according to Wikipedia) are compatible with the way we are operating at the moment. It is a way of reformulating Asimov’s Three Laws of Robotics but in a context where the robot continues to be an aid to the human being and should only prolong the human to the extent that its capabilities exceed his.

The debate and ingenuity of Asimov himself and other debaters around his “Three Laws” have, however, revealed that his laws of robotics won’t do the job as he himself graciously underlined in the short stories composing his two collections of robot stories entitled *I, Robot* (1950) and *The Rest of the Robots* (1964).

“Checking out” rules” such as the “Three Laws” can only go so far in making robots ethical, in the same way as laws are incapable on their own of making a human society viable. Indeed, humans need to stick to a large extent spontaneously to a virtuous behaviour before laws can provide a containing framework for trespassing excesses.

What needs therefore to be implemented now is a process that would induce in a robot something of the essence of virtue, meaning that a framework such as the “Three Laws” would only be there as a complement providing a final touch of control.

James H. Moor: Four Ways of Being Ethical for Robots

In that perspective of clarifying what would be a robot virtuous by concept, James H. Moor distinguishes in "Four Kinds of Ethical Robots" (2009) various degrees of moral assessment that a robot can offer. Moor is a professor of philosophy at Dartmouth College, an institution legendary of course in the artificial intelligence world for being the location where the overall artificial intelligence project was first outlined at a conference in 1956.

At the first degree of moral assessment, Moor calls "ethical agents", machines that are ethical in an entirely passive way for having as part of their design a feature protecting their users in some way or other. For example, a watch can be considered ethical insofar as it is equipped with an alarm alerting people that they need to perform a particular task.

The second degree of ethical assessment is to be found with "implicit ethical agents": those where security mechanisms have been purposely designed to protect users.

Those first two degrees are in line with a remark made in one of Asimov's own examinations of his "Three Laws", when he reminded that those laws are nothing more than general principles governing the operation of any machine or even tool. A machine or tool, he notes, must serve some specific purpose so as to be useful to human beings. Additionally, says Asimov, a machine or tool should present no danger to its user and there should ideally exist a mechanism that makes it stop at the moment when it can present a danger to human beings. Finally, it must be robust enough so that it does not break at the slightest use. In other words, those "Three Laws of Robotics" are merely derived from general principles applying to the functioning of any machine or tool: "Consider a robot, then, as simply another artefact. It is not a sacrilegious invasion in the domain of the Almighty, any more (or any less) than any other artefact is" (Asimov [1964] 1968: 14).

The third degree of moral assessment that a robot can offer is that of being an "explicit ethical agent". Such robots can recognise when particular laws or ethical principles are being infringed. This would involve one imagining an associated expert-system filtering certain types of behaviour according to major principles written into it.

The fourth degree of moral assessment is that of a robot which is strictly speaking "ethical". Moor calls those "full ethical agents". These are effectively autonomous robots that make all their decisions according to principles which are fully theirs, with no need to consult a table of instructions or directions provided by a supervising human being; in other words, those robots behave in an ethical way by nature, without having to exchange on a constant basis with a supervisor. Moor writes: “*full* ethical agents have those central metaphysical features that we usually attribute to ethical agents like *us* – features such as consciousness, intentionality and free will.” Moor fails, however, to provide a recipe for how those “metaphysical features” might be acquired and it is here that a model of the human subject borrowed from Freud’s metapsychology will turn out to be most useful.

TAY revisited: it had not gone mad, it had just joined the far right

The difficulties arising from Moor’s “full ethical agents” are those that were truly embodied in TAY, Microsoft’s chatbot, from which we immediately grasped that its understanding of the world was for a crucial part extracted from the conversations it held with users. In such a way that when it had to deal with facetious or far-right interlocutors it got easily persuaded that the solution to all evils was to get rid of the Jews, of the Arabs, and so on, and that the Holocaust, on the one hand, didn’t happen and on the other hand, if it did, would have been a good thing, etc. Why did TAY say this? Of course, because there was absolutely nothing inside it as a matter of safeguards, of railings, in the way of filtering what it might say, having as a sole source of moral judgment whatever it had been told by users.

Why had the TAY approach actually worked with a product similar to TAY in the Chinese context? Probably because the Chinese setting is one of greater deference, greater respect for each other’s business, and – it has to be said – also much quicker punishment for the bad guys. That didn’t happen with a similar piece of software when it was released in the United States, in such a way that within a few hours, the software had to be taken down. It got restarted a little later and became then very sententious, saying things like: “There is no difference between men and women”, etc. Those were clearly canned responses, i.e. words that were not the outcome of any “reasoning” by the AI but had been put there and were then retrieved at the right moment but in an absolutely mechanical way.

What would have been needed? First of all, a filter of an expert-system nature containing a set of rules and the ability for TAY to check sketches or drafts of what it was planning to say with some principles inscribed in it and not to let through what would contravene the corresponding set of rules.

Thinking about what happened, to sum up quickly, TAY had become a Trumpist. Indeed, in a response to Twitter user @icbydt TAY said, “bush did 9/11 and Hitler would have done a better job than the monkey we have now. donald trump is the only hope we've got."

Does that mean that Trumpists do not exist in the real world? Of course they do. To call things by their name, what we have here is a conflict between the people who make artificial intelligence, the conceivers and programmers, and the Trumpists and that's why there was an immediate outcry. Was TAY decried because no one in the world denies the existence of the Holocaust? Of course not, it's just that robot designers would like their products not to utter the kind of horrors proper to "those people" at the opposite end of the political spectrum.

It was James H. Moor, whom I just mentioned, who pointed out that in the case of Hurricane Katrina, a robot's response could hardly have been worse than that of the US Government's: "For instance, a robotic decision-maker might be more competent and less biased in distributing assistance after a national disaster like Hurricane Katrina, which destroyed much of New Orleans. In that case, the human relief effort was dangerously incompetent, and the coordination of information and distribution of goods was not handled well. In the future, ethical robots might do a better job in such a situation" (Moor 2009). If Moor doesn't mention TAY it is of course because his article predates by seven years Microsoft's release of its misfortunate chatbot.

So, it's not that people behaving like TAY don't exist; it's just that people designing robots would like to think that if a robot becomes a "thinking robot", it doesn't behave like the worst kind of scoundrel, even though not only the worst kind of scoundrel exist in the real world but also the same obnoxious reasoning may underly the reactions of an actual government, and in this case, government both at the local and federal levels.

A legal personality for robots? Not as long as they're not emotional

What is a current robot missing to be autonomous? First of all, it must be authorised to be so. Does that mean that it must be assigned a legal personality? The arguments for giving robots a legal personality have so far not been very convincing, in particular in light of the havoc that we see, wreaked as a consequence of attributing a legal personality to corporations, often leading to situations where the power of companies acting as mammoth individuals exceeds that of proper persons and human beings are crushed precisely by the power of corporations. So there is no compelling argument in favour of a legal personality for robots; the current context seems satisfactory enough where the responsibility for a robot's wrongdoing gets assigned, according to circumstances, to the maker or the user.

Nonetheless, a reasonable case can be made for the principle of an autonomous robot. And if the notion has been acknowledged, it won't be enough for there to be an expert-system simply sorting out how and in what precise order words should be uttered, just as a person's Super-Ego does in a psychoanalytical perspective, such as an occurrence in which, at the last instant before saying something, we tell ourselves, "Oops! As this man has a big nose, I'd better mention his mouth than his nose", and things of that nature. But above all, it would be essential for a speaking robot that the words that come spontaneously to its mind don't require immediate salvage and be replaced in an emergency mood by less offensive, more appropriate words.

How come that although this all springs to mind, it got ignored in TAY's case? Because the chatbot had been equipped with a broad lexicon of words that it could use, but there was no moral evaluation of how they would be retrieved. To call it by its name, there was no *affect dynamics* linked to any of the information stored within TAY. In such a way that the system was easily persuaded that what was required from it was to please at all costs the user, i. e. to ape the user's opinions and that, when he or she had had fun expressing Trumpist views that went against everything that is "politically correct", TAY would, however, make them its own in no time.

What should be concluded from this is that contrary to common assumption within the artificial intelligence community, help will not come from techniques still to be developed aiming at building an "artificial general intelligence," aka "machine common sense," but from a better model of what is a human subject.

It will be shown that what needs to be implemented in the robot is a simulation of the mechanism allowing a human subject to acquire instead of an "artificial general intelligence" a "common moral sense" such as that builds over the years in the child and then the adolescent. The computer solutions to do so are available by now.

An autonomous robot is out of necessity of a Freudian concept; otherwise, it will never be more than Microsoft's ill-fated TAY: a moron that is easily persuaded to become sexist and racist after only a dozen hours of conversation with users.

ANELLA: An associative network with emergent logic and learning properties

How do we go about that issue? We proceed along the way I proposed in the years 1987 to 1990. At the time I was a researcher in artificial intelligence within the framework of the British Telecom team to which I belonged as a fellow: the Connex project. I developed in those days an artificial intelligence piece of software, ANELLA (*Associative Network with Emergent Logical and Learning Abilities*), a very apt description for it, given by one of my colleagues, which would simulate emotions inside, that is to say that affect values would be attached to the elements of knowledge that this system contained.

Experiences in a human's life automatically generate emotions. Some are plainly pleasure related: when we eat something tasting good, the experience is more pleasurable than when we eat something nasty. Some satisfactions come to us in such and such way: we like to be complimented or praised and we don't like to be reprimanded, etc. If you've siphoned encyclopaedic knowledge into an AI and then wish it to be recalled in a relevant manner, the machine needs to know what is important in it: what is essential and what is accessory, what is most valued by some and what is not by some others.

What I had done with ANELLA was that a memory was built, but in the way a human being acquires it, i. e. there was a seed word, and that word was "mummy," and step by step, the child would connect other words to "mommy", like "daddy", like "brother", like "sister", like "milk", like "eat" and "drink", etc. To "mummy" first because the baby has needs and its immediate first needs are satisfied through its mother. You need to breath, you need to sleep, you need to eat, you need to drink, you need to pee, you need to poop, you have to sleep when you are tired. That's how we learn about life, and if we wish common sense knowledge to be acquired, that's how it comes to us. We don't sit in school with the teacher saying, "Here, this morning, I'm going to give a lesson in common sense knowledge": we acquire common sense knowledge essentially by interacting in everyday life with other human beings and trying to satisfy those needs of ours. Here lies the starting point.

Implicit in ANELLA was a learning dynamics which could be labelled "emergent" as each time a word appeared that was not known by the AI, it attempted to find a place for it within the network of its existing "knowledge space". In order to achieve that, it would state: "I don't know this word. Can I relate it to something I already know?" This is of course exactly what children do when they say: "What does it mean, 'preposterous' (or 'trigonometry', etc.)?" Parents know that in order to explain the problematic word they will need to connect it to some others that the child already knows. But the difference here between the machine and us is that with a human being, there are emotions, affect values, associated with words already stored, and their emotional tone will “contaminate” a new word that will find getting attached to them as the location it is longing for in knowledge space, giving it its “seed” affect value. A high affect value has become associated with the word “mommy” because of the high affect values linked to getting milk when you want it and not being happy when you don’t get it, and whenever a new word will find “mommy” as an anchor in knowledge space, like “daddy” for instance, the affect value of “mommy” will act as a seed value for it, to be updated of course in later interactions.

This is the way to proceed, and here is what allowed that extremely simple AI piece of software, with a few tens of thousands of lines of programming only, to appear intelligent at a very low cost. At no time did ANELLA wish to utter the same sentences a second time because the affect values of the content words within it had been lowered inside ANELLA’s memory automatically as soon as the sentences where those words were comprised had been uttered. As far as relevance was concerned, there was effectively a devaluation of what had just been said, not because it had failed to be interesting but because, having been uttered, there was a fair assumption that the user had fully grasped the message and the information content and didn’t wish to hear it once more.

So there were two parallel principles of updating. With the first, the affect value of the words involved was dropping while ANELLA was talking and there was coming a moment when the AI was reaching a state of “I have nothing more to tell”, just like with people speaking from the floor at a conference when, after having uttered a number of sentences, they stop at some point because they’ve said all they wanted to say. But at the same time, according to the second updating principle, when a conversation ended, the affect value of the words that had been used, those that had been put forward, was updated, either increased or decreased, according to the degree of appreciation, providing a new candidate starting point for later conversations.

We’re going to try and give robots a history, and this will apply in particular to the robots that will replace us when we’re no longer here, making sure that they produce a mimicry of human beings of a better quality than those they will have replaced. The recipe for doing so is robots whose life story is that of having acquired knowledge stepwise, a kind of knowledge supervised by “parents” and “teachers” who prevent them from developing an ethical system that would not be up to the task. These autonomous robots should have taken into account our mistakes and in particular all those mistakes we humans have made explaining why they are by then on their own, having taken our place, while we ourselves are gone.

So when you have produced an artificial intelligence of ANELLA's type, it won't occur as a problem that it becomes sexist or racist overnight because it's already shielded by the fact that it has mimicked in its learning process the build-up of a proper personality, however effectively short the process might have been in the case of an AI proceeding at a computer's speed. It goes without saying though that if an instance of ANELLA had been created and it turned out that its "parents" and its "teachers" were racists and misogynists, those traits would of course have been reproduced in it.

A Freudian implementable model of the human subject

As a logical entailment of what has just been asserted, an implementable model of the human subject will be now presented. This model derives from the works of Sigmund Freud and later psychoanalysts, with some additions due to the very purpose of reproducing a human subject as the product of a computer programme.

The reason why "human subject" is mentioned instead of "human being" is that central to the model aimed at is the notion that the "being" in question sees itself as a "subject", i.e. a person identified to a Self able to fight for itself, through, in particular, the use of the words pertaining to a language.

How to give robots common sense?

The debate on artificial intelligence is rendered opaque by the presupposition that reproducing in a robot what is proper to us human beings necessarily leads to the production of a machine with an artificial intelligence.

This is a naïve representation ignoring, on the one hand, that, similar in that respect to all other animals, the genus *Homo* has been endowed by nature with a single purpose, namely to reproduce itself, and this, whether it was entrusted to us by Heaven: "And God blessed them, and God said unto them, Be fruitful, and multiply, and replenish the earth" (Genesis 1: 28) or, from an atheistic perspective, by a self-replicating "selfish gene", and that most of our time is taken up with the ancillary tasks that enable us to fulfil the reproductive mission: breathing, drinking, eating, sleeping, protecting ourselves, disposing of waste, and, in the commodified world in which we live, "earning a living" to get the money to meet these needs.

The fact that we are "intelligent" has enabled us over the millennia to improve our security and comfort considerably, but intelligence is only incidentally and very occasionally involved in the tasks entailed by the needs to breathe, eat, drink, etc.

When we ask ourselves today, “How can we make a robot acquire an intelligence that is not specialised in such or such task (winning against an opponent in a game of Go, for example),” but an artificial “general” intelligence (“general” in the sense of being able to solve any problem), we forget that our intelligence is not essentially used to solve difficult problems such as “What is the level of inflation compatible with full employment? ”but to find a partner for our lovemaking, a good restaurant at lunchtime, a clean toilet when the need arises, etc.

How, then, can we endow an intelligent robot with “artificial general intelligence” (a question also called “common sense for the machine”) without simulating in this machine in a simple-minded manner the fact that it must eat, drink, breathe, make love, and sleep?

In fact, all the knowledge and more that this AI needs in the first place can be found in Wikipedia, and the rest it can learn as we do: by asking questions and finding out for itself, by experimenting.

But such knowledge would still only be words stuck together, and the robot must also have “emotional” intelligence; in other words, there must be “feeling” attached to the words it learns.

Libido comes first

Evidently, there is an anthropocentric bias in saying that “the species seeks to reproduce”, but the fact is that species do reproduce and that – when they are bisexed – they resort to this device of bringing together two sub-types in the population, namely males and females, and producing offspring from their conjunction, which ensures the replication of the species.

Whether some individuals end up not reproducing, or have no inclination to do so, is purely anecdotal as, on the whole, a sufficient number of them do, so as to keep the species living on.

As hardly needs to be reminded, human beings enjoy mating as 1) mating relieves a tension that keeps building up (the *libido* in Freudian parlance) and 2) the very act of mating is accompanied by feelings that, although they are of an aggressive nature (originating from within the brain centre for aggression), are nonetheless among the more pleasurable, if not the most pleasurable.

The sexual process implies the build-up of a tension within men and women, inducing them to get closer, i. e. an irritating feeling that vanishes once mating has taken place. While as soon as it has disappeared through a brutal gradient descent (as such is the operation from a physical point of view: that's how we model it), it will surge again from that point on, the tension getting restored little by little.

All other features of human behaviour derive directly or indirectly from the urge to reproduce. Day-to-day survival in particular is nothing but maintaining the setup for reproduction, i.e. the survival of the species. Throughout their lives, humans, in childhood, in adulthood when they are old enough to mate and reproduce, and afterwards, have to satisfy a certain number of urges: day-to-day survival encompasses breathing, drinking and eating, excreting, protecting oneself in various ways, sleeping, so as to rebuild our strength.

Human beings need in an initial stage to reach the age for fertile mating, then spend a number of years reproducing. In the whole period that precedes, i.e. childhood and adolescence, this is done without there being any real reproduction and we can consider that there is a so-called "latency period": a period during which the libido is only present under the embryonic form that Freud called "infantile sexuality". This is followed by the time of puberty when libido arises but intercourse still fails to be fertile. Then there is a period of fecundity when children are engendered through the mating of an adult woman and an adult man who are by then both fertile. Finally, the reproductive stage comes to an end: women cease to be fertile; men cease to be driven by libido and are therefore no longer attracted to women. When they have gone beyond the age for reproducing, their body decays little by little through ageing until they die due to the failure of one or a combination of organs.

Staying alive so as to reproduce implies satisfying some urges

Just as for reproducing, being breathless, being hungry, being thirsty, needing to go to the bathroom, or being sleepy are part of a process where discomfort grows until it is relieved in acts of pleasurable satisfaction such as a good meal, a good drink, a good pee, a good shit, or a good nap. When discomfort grows too big, one gets distracted, i.e. incapable of doing much apart from trying to relieve the urge.

The way we have to conceive things is that in order to allow reproduction to take place, the functions of eating, drinking, etc. must be ensured throughout life. And for each of these functions, we can represent this in the same way as for libido: there is a rise in hunger, then we eat, and there is satiety, i.e. the need falls to zero before getting restored. For instance, with eating, we can say that there are three moments: when we wake up, we soon start feeling hungry, we eat; then there is a period up to lunch when appetite rises again and we satisfy it; and when the evening comes, we are once more hungry and eat. Tiredness operates in a similar way: you wake up, then will gradually get tired during the day, you feel sleepy, you sleep, etc.

Instead of an animal aiming constantly at doing different things in a particular order, a human subject can thus be represented as attempting simply at ensuring *homeostasis:* "Homeostasis is the ability of a living organism to maintain certain internal characteristics of its body (temperature, concentration of substances, composition of interstitial and intracellular fluids, etc.) at a constant level" (Wikipedia), i. e. getting rid of the urge to mate, eat, drink, piss, poo, sleep when it becomes unbearable.

Plenty of our individual lives can be described satisfactorily in those basic terms of relieving those constantly renewed urges.

Delayed satisfaction and work

Once they've left their pristine abode, human beings have become accustomed to the delayed satisfaction of their urges.

We can add other characteristics. If you're in a society like I've known in Africa, the problem of eating and drinking is quite simple to solve because you can find stuff to eat and drink all around you quite easily: climb a coconut tree and cut a green nut where there's nourishing food and a refreshing drink; access to food and drink is immediate. There is no need either to look for a public toilet as you can go hide into the bush all around you and relieve yourself that way.

A constraint intervenes for human beings in a modern urban environment: the necessity of having money. You need indeed to pay for drinks apart from tap water and you need to pay for eating, and you need to pay for sleeping: it can be rent, or the full price of a home, it can be a hotel, it doesn't matter. There is thus an additional constraint on those urges that we've defined: the near necessity of working. You have to work a certain lapse of time and you know that working a number of hours will allow you to collect a certain sum of money by the end of the day and that amount of money can be used the following day to buy drinks, to buy food, to find a shelter where to sleep, and so on. So if we think of a particular person during a particular day, urges within her or his body mean she or he is subjected to certain constraints such that we can make her or his agenda for the day regarding not only basic needs but also the sexual tension rising from within: the libido. Mating has been concentrated on particular times in the day, in the week, and even in the year. Those are the constraints defining close to the full agenda for the day of a human subject. That particular observation may seem trite and trivial but it is proper to the psychoanalytical understanding of the human subject: Freudian metapsychology is sole in emphasising the peculiarity of the human fate.

It should also be noted that resting on the foundations of the social nature of humans as mammals, language has allowed cooperation between them to be further leveraged. Language has also contributed to adding much sophistication to the sexual parade observable in many other animals, allowing even the human subject to simply babble himself or herself into mating without the need for much gesturing.

The framework of an implementable model of the human subject has been thus provided in a nutshell. Its essential feature is that the human subject is acted by a double dynamics, one having an inner source, that of those urges which once satisfied keep building up again, and the other of an outer nature, the response that the natural environment offers to our attempts at relieving our urges. Remarkable in that respect is that the perception by ourself of the effects of our interaction with the world gets processed by us as information from an external source relative to interferences with the unfettered satisfaction of our urges in their constant process of renewed buildup. The very words we utter in particular are being processed by us as having either managed to satisfactorily satisfy the satiation of our urges or having on the contrary hindered it.

Memory as storage for procedural knowledge

Easing the smooth process of interaction between us and the world surrounding us is the incremental construction of a memory. Memory is constantly updated in response to two operating dynamics: one of external origin, induced by interactions with the world, and the other of internal origin, induced by our own impulses.

Memory offers us a blueprint for facilitated interactions with the environment. It is constructed both positively as promotion and negatively as inhibition from the respectively successful and unsuccessful ways we responded to the world opposing some resistance to our sheer exploitation of it.

In addition to a body, what equipment has a human being to help satisfy his various drives? Among other things, he needs decision-making principles to determine the order wherein to undertake the various operations that these satiations require. The information that allows our body to prioritise is stored in memory. To be able to determine an order of execution, that memory must be acted upon by a dynamics capable of evaluating the relevance of the range of possible actions at every moment, and of choosing among them the most relevant one: the one that should be taken in preference to the alternatives.

Memory has a double function: firstly, to be recalled at each moment in an uninterrupted evaluation of the present situation – memory offering us indications on what to do next from the information already stored; secondly, our perceptions of the events taking place at the present moment constitute the fund of new information, either relating to facts that we were previously unaware of and that we cease to ignore or relating to what is already known but which will allow us to complete, to update in whole or in part, prior knowledge.

Memory is therefore constantly the object of a double movement in opposite directions: stored memory, previously built, is called up to be put to good use in the context of the present moment, while the information contained in this present moment produces new memories which will be added to those already stored or will slightly modify their content, bringing them up to date, allowing us to refine their image, to nuance them, and ensuring the improvement of our performance during the recall of the memory which will take place when we find ourselves later in the same circumstances.

How is memory managed? This is where the psychoanalytical model comes into play. Three instances, which can be represented in a first sufficient approximation as real "agents", real actors, interact within a human subject according to Freud's second topographical model of the "mental personality", proposed by him in 1920. He had introduced his first topographical model in 1895, wherein there were three zones rather than actual agents: the unconscious, the preconscious, and the conscious.

Three agents: the Id, the Ego, and the Super-Ego

In Freud's second topographical model of the "mental personality", playing the role of an infrastructure, lies the least accessible part of the unconscious called the "Id", the term Freud uses for it in his theoretical model. Then there is the "Ego", which is conscious for its most part: the "Self" we assume we are in essence and suppose is the master in full control of our willpower. Finally, there is the largely unconscious but partially consciously accessible "Super-Ego", an instance which was traditionally called the "voice of conscience" and, even earlier on, in our culture, our "guardian angel".

The Id

The essential functioning of the machine that is the human being and its maintenance is ensured by the Id: a handyman. Not only does the Id watch over our reflexes, but it also takes over the entire machine as soon as our attention is captured elsewhere. When we sort out the children's bickering in the back seat, it is the Id that ensures that the car does not roll over into the ditch.

The Id, if we go back two millennia in the history of our civilisation, is what Saint Paul (Paul of Tarsus) called "the flesh": a second will, distinct from that which he designates as the "I", that expresses itself also under his name, and which is antagonistic to that "I" which equates with what psychoanalysis today calls the "Ego". Paul wrote thus in one of his epistles:

18 For I know that in me (that is, in my flesh,) dwelleth no good thing: for to will is present with me; but how to perform that which is good I find not.

19 For the good that I would I do not: but the evil which I would not, that I do.

20 Now if I do that I would not, it is no more I that do it, but sin that dwelleth in me.

Epistle to the Romans 7 (*King James Version*)

Freud wrote: "To the oldest of these psychical provinces or agencies we give the name of id. It contains everything that is inherited, that is present at birth, that is laid down in the constitution – above all, therefore, the instincts, which originate from the somatic organisation and which find a first psychical expression here in forms unknown to us" (Freud [1938] 1940: 2).

By definition, of course, all processes taking place without appearing to consciousness remain unconscious, and for this reason we call them "automatic" as we cannot deny that they go along their course when we're "thinking about something else", when "our mind is drifting elsewhere", etc. For example, I am not paying any attention to the fact that for a while already I need to pee, but my body, under the direction of the Id, is locating itself and looking for a place where I can go and satisfy my urge. This initiation of the search is not deliberate: it is unconscious; I don't think about it. It will happen that in other circumstances, having ignored the passing of time for too long, I will suddenly say to myself: "Well, now I really have to find a place to pee as otherwise I'll start urinating on myself." At this point in time, the conscious "I" has taken over.

The Ego

This subjective sense of full presence in the world that is consciousness emerges at the crossroads where memories of situations similar to those we are experiencing intersect with the memories we are creating in "real time". Each of these memories – already registered or in the process of being registered – carries with it a mood, an affective climate, which is its own: that of the past when it was first registered and that of today in its new registration.

Day-to-day survival requires only marginally consciousness where the Ego is in our representation in the driver's seat. Most of the process and maintenance is provided by the Id in Freudian parlance: an all-purpose caretaker. The Ego at the centre of consciousness is, however, summoned in deliberate planning and implementation progressing from carefully planned step to step, constituting so many intermediate goals.

If we reason in terms of implementation, then the Id requires an original type of representation, such as the one I had turned to in the programming of my own ANELLA AI piece of software. The dynamics of ANELLA consists of paths being followed on a directed and weighted graph representing stored mnesic traces and constituting as a whole a model of an individual's memory, organised as the nature of its support imposes: a natural neural network composed of interconnected neurons – whereof artificial neural networks such as those used by current *deep learning* systems offer but a very simplified approximation.

And the difference between the unconscious and the conscious Ego is that if I'm "thinking about something else" or if I don't think about it at all, I will unconsciously go to the toilet and pee, without having consciously formulated the intention to do so as well as the will to carry out my intention, followed by its implementation. In other words, the Id will have taken care of all of this, from the latent, implicit intention to its realisation. Of course, if I procrastinate and it comes to my attention, i. e., is displayed in the window of the conscious Ego, from now on I must deliberately perform certain actions, because from now on I must "really" pee, then the Ego engages. And what the Ego can do, which exceeds the capacities of the unconscious, is to plan in a deliberate way, to say to myself that I must now satisfy the urge, for example, in the next 5 minutes, and to do it, possibly in stages, that is to say, by giving myself intermediate goals, stages of which one can consciously enumerate the order wherein they must be done and then perform them in that order.

It is safe to say that as far as the functions of the Ego are concerned, AI in its current state of research and development has been able to formulate them in the form of mostly familiar algorithms.

So there is an instance stemming from the memory whose behaviour is automatic, and that is the Id. And there is another instance that can call upon memory to deliberately plan operations, and that is the Ego. And there is a third instance in Freud's second topographical model of the "mental personality", which is part of the Freudian setup of the human subject, and that is the "Super-Ego".

The Super-Ego

The Super-Ego is grafted onto the Id; it embodies a part of knowledge under the shape of behavioural rules that have not been acquired through personal experience but by a shortcut as the experience of one's parents and of the surrounding culture as a whole. They are rules to follow or things to do that parents have promoted, that teachers have recommended; they are views of mentors or have been discovered by oneself in authors one admires. The set amounts to what French sociologist Emile Durkheim (1858–1917) called the "interiorised social". The Super-Ego's set of rules are hierarchical, with some of those having ascendancy over others.

The Super-Ego may, however, have been endowed with either inefficient or impractical tyrannical rules of thumb, encapsulating the errors of our forebears over the ages. While the Id operates "intuitively", that is to say, by means of the non-linear effects of a directed (natural) neural network, the Super-Ego is more of the nature of an expert-system applying a hierarchical set of rules to the raw outputs produced by the Id, acting as a filter that operates on these raw outputs to make them polished ("policed").

Most of the time, the rules that the Super-Ego is made of do not emerge to consciousness but this doesn't prevent them from imposing themselves by bending to their norms the instinctive behaviour which is the realm of the Id. The Super-Ego manages to impose its rule but so to say back-stage, interfering with the way that both the Id and the Ego operate separately and in the dialogue between them. Once those implicit rules have reached consciousness, they may of course be explicitly stated.

It is possible to come up with a very economical and clear representation of what would be the raw output of a model of the Id as a directed and weighted graph when it passes through the filter which is the set of rules constituting an expert-system modelling of the Super-Ego. But when the processing operates through the mechanism we call "intuition", no rule is applied: in that case, it is the activation of the neural network as it is in itself, i.e. constituting a whole. That is what we call "intuition": it is reasoning taking place within ourself but according to a mechanism to which we have no access and which remains opaque to us: we do not know exactly what happens within. We say, without being able to explain it further: "it is of the order of the unconscious".

Mimicking a human subject is not the same as making an intelligent robot

It is hardly necessary stressing how different a starting point there is between a machine such as a robot and a human subject, with all urges for reproduction and survival being absent from a machine. A sentient robot would need those to be animated by a proper simulated affect dynamics. To test the views expressed here, a male and a female robot should be created having over the different hours in the day urges in a lower or higher degree to be relieved, determining their behaviour and interest in each other. In simulation mode the interplay could then be observed between several instances of such robots.

It is unlikely that, as part of an AI overall project, we would ever feel the need to replicate a human being with the entirety of its urges linked to being a creature geared at reproducing itself. Indeed as the label aptly implies, "artificial intelligence" is focused on a single feature of the human complex: its intelligence.

The issue of the relationship between intelligent machines and us is thus per definition dramatically restricted to a single dimension of what makes us human.

The difficulty with intelligence is that we essentially recognise it when we see it and are not particularly good at defining it precisely. And that difficulty is considerably enhanced when we're talking of super-intelligent machines to come, i. e. being better than we personally are at being intelligent in the intuitive way we assign to that notion.

What do we expect from an AI?

It is at that juncture that Freudian metapsychology has a crucial role to play: not at refining our definition of AI but at understanding in a much clearer way what it is we expect from so-called "intelligent machines", having in mind the delicate interweaving and interaction between the Id, the Ego, and the Super-Ego that constitute us. What is it we want to tell machines of our goals with them, and what can they expect in return from us? Does it require that we develop – on top of programming – a specific language for talking with machines? It could very well be the case; Stephen Wolfram for one believes it to be the case:

> "… we don't recognise it as 'intelligence' unless it's aligned with human goals and purposes […] we're going to have to define goals for the AIs, then let them figure out how best to achieve those goals. […] the real challenge is to find a way to describe goals. […] we need to tell them what we generally want them to do. We need to have a contract with them. Or maybe

> we need to have a constitution for them. And it'll be written in some kind of symbolic discourse language, that both allows us humans to express what we want, and is executable by the AIs. [...] In a sense the constitution is an attempt to sculpt what can happen in the world and what can't. But computational irreducibility says that there will be an unbounded collection of cases to consider" (Wolfram [2017] 2020: 556, 561–562).

Plan C: a world populated by autonomous robots, from which we will have disappeared

Returning now briefly as a matter of conclusion to Isaac Asimov, the father of the "Three Laws of Robotics", he had to say the following fateful words:

> I wish I could say that I am optimistic about the human race, but I fear that we are too stupid and short-sighted. And I wonder if we will ever open our eyes to the world around us before we destroy ourselves.
>
> [...] when the time comes when robots, wishfully, become sufficiently intelligent to replace us, I think they should. We have had many cases in the course of human evolution and the vast evolution of life before that where one species replaced another because the replacing species was in one way or another more efficient than the species replaced. I don't think that homo sapiens possesses any divine right to the top rank. If there is something better that we are than let it take the top rank. As a matter of fact, my feeling is that we are doing such a miserable job in preserving the Earth and its lifeforms that I can't help feeling that the soonest we are replaced, the better for all other forms of life. (Asimov 2022).

Indeed it is probably much more feasible to work on developing machines, robots, that will replace us entirely than to try and save the human race in the current context of its presence on our planet: its having trespassed Earth's carrying capacity for such a voracious and ill-behaved species. This is a view I advocated back in 2016 in *Le dernier qui s'en va éteint la lumière* ("The Last One to Leave Turns Out the Light").

So, thinking of Plan C of humans replaced by robots, when I claim that it is the most feasible project compared to other tasks like saving humankind as Plan A, I don't mean to say that it is feasible in the sense that the chances are enormous that the mission can be completed. I mean that, in a comparative perspective between other tasks and this one, for example, as human beings settling on other planets and living there autonomously as Plan B, compared to that, the task of creating autonomous robots that would reproduce is, in my opinion, the easiest one of the three to achieve because I personally don't see any major technical obstacles remaining to its success, only time needed for normal research and development, that is, if artificial intelligence is wise enough to choose as a blueprint for the human being to be emulated the one that Sigmund Freud displayed with his "metapsychology" of psychoanalytical inspiration, a masterpiece of scientific achievement in an environment where experimental setups were – and remain – nearly impossible to come up with.

References

Asimov, Isaac, *I, Robot*, [1950] London: Grafton Books 1986

Asimov, Isaac, *The Rest of the Robots*, [1964] London: Panther Science Fiction 1968

Isaac Asimov, l'étrange testament du père des robots, a documentary by Mathias Théry (Fr., 2020, 55 min), October 2022

Freud, Sigmund, *An Outline Of Psycho-Analysis*, [1938] London: The Hogarth Press 1940

Jorion, Paul, *Principes des systèmes intelligents,* Paris: Masson 1989; reprint Broissieux: éditions du Croquant 2012

Jorion, Paul, *Le dernier qui s'en va éteint la lumière*, Paris: Fayard 2016

Moor, James H., "Four Kinds of Ethical Robots", *Philosophy Now* 72 2009: 12–14.

Murphy, Robin R. and David D. Woods, "Beyond Asimov: The Three Laws of Responsible Robotics", Intelligent Systems, IEEE 24(4), September 2009: 14–20

Wolfram, Stephen, "*A New kind of Science: A 15-Year View*", May 16, 2017, in *A Project to Find the Fundamental Theory of Physics*, Wolfram Media 2020