

# Cognitive automata and the law\*

Giovanni Sartor

December 16, 2003

## 1 Forward

Our technological society is populated by more and more complex artificial entities, which exhibit a flexible and multiform behaviour. Such entities increasingly participate in legally relevant activities, and in particular in negotiation. Already today many contracts are made by computer systems, without any human review. In particular, this happens through software agents (SAs), who may execute autonomously the mandate that has been assigned them, without any subsequent contact with their human users.

In this context, we naturally tend to apply also to artificial entities, and especially to SAs, those interpretative models which we apply to humans. In particular, we tend to explain the behaviour of such entities by attributing them mental states (beliefs, desires, intentions . . .). Consequently, we tend to qualify their actions through legal notions which presuppose the attribution of mental states. Consider, for example, the possibility that a computer system enters into a contract (wants to make the contract, and to realise the legal results it states), is the object of a fraud (is cheated), makes a mistake (has a false belief), harms somebody with malice (intentionally), etc.

Our natural tendency to attribute mental states to artificial systems, and to apply the consequent legal qualifications, can be contrasted according to the view that all mentalistic concepts only apply to humans. This thesis would imply the necessity to undertake an extensive review of the existing legal notions, in order to apply them also to the relation with or between

---

\*Published in Bing, J. and G. Sartor. 2003. The Law of Electronic Agents, 67-114. Oslo: Unipubskriftserier

artificial systems. We would need to eliminate from such concepts any connection with mental or spiritual attitudes: the element of will needs to be eliminated from the notion of a contract, the element of having a false belief from the notion of a mistake, the element of producing such false belief from the idea of misrepresentation, the element of intention from the notion of malice . . .

However, it is dubious that this strategy may lead us to results which are appropriate to the needs of our time. In fact, it forces the legislator and the jurist to make the following choice: either to eliminate any psychological notion from the law, or to duplicate the characterisation of legally relevant facts, providing besides a mentalistic characterisation, to be applied to humans, a purely behaviouristic characterisation, to be applied to artificial entities. Neither of the two options is very appealing. The first produces not only a conflict between legal qualifications and the usual interpretation of social facts, but also a clash between legal evaluations and our intuitive sense of justice (according to which the mental states which have determined and accompanied an action are a decisive element for its evaluation). The second option requires not only duplicating the legal systems, but also maintaining the consistence and coherence between mentalistic and behaviouristic norms.

In this contribution we will state that this choice does not exhaust the available options. It is possible to interpret mental concepts in a flexible and neutral way, so that they become applicable also to some types of artificial entities. This would allow us to preserve both the spirituality and the unity of the law, even in a society which is increasingly characterised by automated information processing. Before illustrating this idea, let us first clear the field from three possible mistakes.

The first mistake consists in assuming that the law necessarily needs to adopt a behaviouristic perspective, since mental states cannot be perceived, and therefore cannot be objectively ascertained from an impartial observer. The fact that the judge (and, more generally, any third party) cannot have direct access to another's minds, would imply the impossibility that mental states are viewed as legally relevant elements, i.e. as states of affairs which contribute to producing legal effects. This reasoning is certainly fallacious: each one of us (and in particular the professional psychologist) has direct access only to the behaviour of other people. However, it is certainly possible (and perfectly legitimate) to infer from such behaviour the presence of certain mental states. The difference between a behaviouristic and a mentalistic approach does not concern accepting different cognitive inputs (behavioural

inputs rather than mental inputs): the difference consists in that the first approach only registers other people's behaviour, while the second, on the basis of behaviour, attributes mental states. The second approach views behaviour (having bought a gun) as providing clues for the presence of corresponding mental states (e.g. the intention to kill). It is undoubtable that the law frequently adopts a mentalistic perspective: when the law uses mental notions (belief, will, intention, etc.) in characterising a type of fact, the judge and the lawyer cannot limit themselves to register observable behaviour (the fact that the behaviour of a person caused the death of another), but the need to consider whether observable facts provide sufficient clues to establish certain mental attitudes (the intention to kill).

The second mistake consists in assuming that attributing mental states to artificial system implies assimilating such systems to humans. According to this view, by attributing significance to the mental states to artificial systems one would refuse to accept that only humans have interests deserving legal protection, that only humans are, as Kant said, ends in themselves. However, attributing legal relevance to the mental states of artificial entities does not imply attributing normative positions to them, in order to protect their own interests. On the contrary we will only examine here whether an intentional states of an artifact can be an element of a precondition which produces legal positions (rights and duties) on the head of natural or legal persons, to protect the interest of the latter (consider for example the case where a person acquires goods or services through a contract which is made through a SA).

The third mistake consists in assuming that a legal analysis of mental concepts presupposes a complete and uncontroversial theory of mind and conscience, which can substitute the intuitions of common sense. If this was the case, we would need to face an undertaking which is not only very difficult, but also legally impossible. The task of the jurist is not that of discovering the scientific truth, but rather that of elaborating proposals which may be successful in the practice, i.e. that of providing lawyers with the chance of converging into shared models. And one cannot expect lawyers to converge into adopting the only best theory of mind, i.e. to succeed where philosophers, psychologists and neurologists have so far failed<sup>1</sup>. On

---

<sup>1</sup>For a collection of some important contributions, cf. Cummins and Dellarosa-Cummins ([CDC00]; for a basic introduction to philosophy of mind, cf. Davies ([Dav98b]). As it is well known contemporary philosophy of mind is characterised by the opposition of competing approaches. Some repropose the so called Cartesian dualism, that is the idea

the contrary, it is well known that when one submits mental concepts to philosophical and scientific examination, the certainties of common sense (which are the starting point for the lawyer) dissolve, and are substituted by complex and conflicting theories. So, the lawyer should not try to propose a new theory of mind, nor to endorse one such theories to the exclusion of all the others, but more modestly, to articulate a point of view which allows one to make reasonable assertions concerning mental attitudes, while respecting common sense.

The starting point for our discussion will be taken from the work of Daniel Dennett, one of the most interesting philosophers of our time, who produced contributions ranging from the theory of mind, to evolution, to artificial intelligence (see, in particular, Dennett [Den89], [Den91], [Den96], [Den97]) . According to this author, we may look at the entities with which we interact according to three different stances, the physical stance, the design stance and the intentional stance (see, for a synthetical characterisation of

---

that mind and brain are different (though connected) realities (see, for all Popper and Eccles([PE77], 355 ff), while others, to use the famous expression by Gilbert Ryle ([Ryl49]). Some believe that the mind is a semantic machine, i.e. that is fundamental capacity consists in formal calculations, as those performed by computers (see Haugeland [Hau00], while assume that the brain in an evolutionary and selective structure, very different from computer systems (Edelman and Tononi [ET00], 212 ss). Some view intentionality and meaning as only pertaining to human mind and language (Searle [Sea90]), others attribute intentionality and meaning also to biological or mechanical processes (Dennett [Den89], Millikan [Mil01]). Some adopt so called functionalist models, according to which the mind results from independent computational processes (Fodor [Fod83], other prefer a holistic vision of the brain, according to which no process can be considered in isolation. Some support the so called eliminative materialism, according to which the common sense notions used in psychology (folk psychology), such as those of belief, intention, or will only are deceptive hypostases, which should be substituted by the concepts produced by or neuro-science (Churchland [Chu00], 500-512), while others believe that such notions are an essential components of individual and social psychology.

There is an even more diverse debate concerning the notion of conscience and the possibility that it can be attributed to artificial objects. Some believe that conscience is fully reducible to the physical-chemical operations of our brain (Churchland [Chu95]). Others assume that it is possible to explain consciousness through a scientific description of its function, and admit that in principle such function can be executed by artificial devices (Dennett [Den91], 21 ff). Other, believe that consciousness can only be captured from an internal perspective, which is unaccessible to scientific objectivity (Nagel [Nag74]). Others, finally, describe it as a physical process with in embodied in a biological individual, which can be scientifically described, though this description will not coincide with the personal experience of the conscious individual (Edelman and Tononi [ET00], 207 ff.).

this distinction see Dennett [Den97], p. 28 ff.). In the following pages we will consider how these three perspectives can be applied to artificial entities, and what legal consequences this determines.

## 2 The physical stance

When we adopt the physical stance, we explain the behaviour of an object according to its physical conditions and the laws of nature that apply to such conditions.

So, we may explain the behaviour of a falling object on the basis of our knowledge of its conditions and of the physical laws of motion. For example, knowing that the acceleration of gravity is about 10 meter per second, I may conclude that the stone that I, emulating Galileo de Galilei, have thrown from Pisa's leaning tower will have, after 0.5 seconds, the speed of 5 meters per second. In the same way, I may explain the shock that I had when I tried to insert my computer's plug into a defective outlet, according to the hypothesis that I touched both the positive and the negative wires, so that electricity ran through my hand. Similarly I can explain why a car went off the road according to the hypothesis that it went too quickly, so that centrifugal acceleration prevailed over friction over the road bed.

The physical stance can be applied not only to an inanimate natural object, but also to artifacts, animals or humans. If I let a stone or a computer fall from the top of the leaning tower, or if I myself am pushed into the air, I can provide the same explanation (of forecast) of the speed of the falling object when it hits the ground, according to the same physical laws. In fact, every object (be it natural, artificial, mechanical, electronic or biological, living or inanimate) obeys the same physical laws, and its behaviour can in principle be explained according to the same laws.

Such explanation, however, though possible in theory, is practically feasible only to the extent in which we know both physical laws and the conditions for their application: if I do not know the general laws of mechanics, or the particular conditions of certain bodies, I will not be able to precisely forecast their movements.

### 3 The design stance

The design stance can be adopted with regard to two types of entities: artefacts and biological organisms.

Let us first consider artefacts. How can I forecast that when I push the button on the back of my piezoelectric lighter it will emit a sparkle? Certainly not because I know the internal structure of my lighter and the physical laws governing it, but because I believe that the lighter has been designed for that purpose (producing sparkles). I assume that the lighter's design is good enough and that it has been implemented well enough: therefore I expect that the functioning of the lighter will achieve that purpose.

In general, when I look at an artifact (a toaster, an automobile, a computer, etc.) from the design stance, I assume that the artifact, having been designed to perform certain functions will really work in such a way as to achieve such functions (it is it used in the way which was intended by the designer). Moreover, I will explain the presence of certain components (in the case of the computer, the screen, the keyboard, the memory cards, the processor, etc.) assuming that these components have certain functions in the design of the artifact and therefore contribute, according to such functions, to the working of the artifact, in the way which was intended by the designer.

So, I can explain the behaviour of my computer (its capacity to receive data, to process them according to a program, and to output the results), according to a functional analysis which distinguishes different components (the so called von Neumann model): an input device (the keyboard), a memory unit, which registers data and instructions, a processor, which executes the instructions, an output device (like the screen). The functional analysis can be progressively deepened. For example, I may explain the functioning of the processor through the fact that it is composed of two components: the control unit, which indicates the next instruction to be executed together with its operands, and the arithmetical-logical unit, which executes the indicated instruction. In the same way, the functioning of each one of the two components will be explainable on the basis of the functions which are executed by their subcomponents, like, for the arithmetical-logical unit, the registers which store instructions, operands and results, the the circuits carrying out the various instruction the processor is able to execute. Such circuits, in their turn consist of logical gates, having the function of executing the basic operations of boolean algebra, and finally, logical gates will

result from combinations of transistors, which let electric energy pass under certain conditions. At this point, to explain the working of the transistors, I need to abandon the design stance, and adopt the physical stance: the working of the transistors will be explainable according to the physical laws of electromagnetism (laws of Coulomb, Ohm, and so on).

In the end, my portable computer will appear to be a physical object, the behaviour of which is in principle fully explainable/foreseeable according to physical laws. This does not mean, however, that the design stance is useless, and that the only way to approach my computer is to explain its behaviour on the basis of the study of electric energy and magnetisation. An analysis at the physical level, though possible in principle, is not concretely practicable towards a complex object, like a computer. The mathematical calculations to be executed are so difficult that they may be unfeasible even for the few people having a knowledge of mathematics and physics which suffices to explain the functioning of all hardware components. Moreover, even those people cannot possibly know all condition which are relevant in applying all physical laws (the electromagnetic condition of every single component of the computer).

In conclusion, when we are interesting in the macro-behaviour of a complex artifact, we have no alternative to the design stance, developed at the appropriate level of abstraction. This does not mean that the design stance is infallible. It grounds expectations that can be proved wrong by reality: the designer, or the implementer of the project, can have made mistakes, and consequently the behaviour of the object can be different from the behaviour the designer wanted to obtain. For example, the circuits which realised a particular operation, such as floating point multiplication, can have been wrongly implemented so that, for a particular combination of input numbers, they provide an incorrect result.

An artifact can also go through degeneration. For example, the structure of my portable computer, as a consequence of various events (an excessive inflow of electric energy, a blow, etc.) can become different from the original structure, and this variation may lead to a behaviour which is different from what was intended.

Finding out that the working of an object is defective means that the design stance, applied to the whole object, has failed. To explain the anomalous behaviour, we need to move to a lower level.

In some cases, the lower level can still involve the design stance, which is now applied to the single components of the object: we may understand

why the behaviour of the object is different from what the designer intended, by finding out the functions of the components of the object, and the ways in which such functions interact. For example, the anomalous behaviour of a program can be explained on the basis of a programming mistake, that is the fact that the program's instructions, each of which works perfectly, have been wrongly chosen or combined (for example, the programmer has written an addition instruction rather than a multiplication one, or has inverted the order of the instructions).

In other cases, on the contrary, the explanation of the behaviour of the system requires moving to the physical stance. For example, the fact that a portion of the screen of my laptop is blank can be explained through the hypothesis that, consequently to a fall, the electric connection which activated that portion of the screen have come off.

In any case, the lower level explanation can integrate, but not substitute the design stance. This explanation can report exceptions with regard to the results of the design stance, which remains the best approach to foresee the normal behaviour of the object, and to identify when it does not work properly. Moreover, the design stance is usually the only approach which is available to us before malfunctioning has taken place.

The kind of functional analysis which is appropriate to the design stance is applicable not only to artificial entities (which are built according to the design of their creators), but also to biological entities. Its application to biological organisms can be grounded upon the assumption that such organisms result from the act of creation of a divinity, or anyway from the plan drawn by a designer (for example, a genetic engineer). However, it may also be independent from such assumptions. In fact, it is possible to assume that the mechanisms of darwinian evolution realise an imperfect, but continuous alignment between each species, and the specific way in which it (the individuals belonging to it) realise the fundamental functions of survival and reproduction<sup>2</sup>

Firstly, only the organisms who could survive and reproduce transmit their genes (and therefore the phenotypic properties which determined their success) to their successors.

Secondly, as far as mistakes (casual variations) in the reproduction of the genetic heritage are concerned, one must keep in mind that those variations

---

<sup>2</sup>We apologise for this trivialisation of the complex problem of evolution. For a "philosophical" discussion of this problem, see Dawkins [Daw89], Dennett [Den96].

which favour survival and reproduction tend to be transmitted to a higher number of successors. Consequently the world would tend to be populated by the organisms which are most able in surviving and reproducing, in the different available biological niches.

Thirdly, an organism may explicate its basic function (survival and reproduction, only if its organs contribute appropriately to this function: if one of such organs did not work properly (e.g., if the lungs could not absorb oxygen), the organism could not survive, and therefore transmit its features. Therefore, while keeping the alignment between an organism and its fitness (its ability to survive and reproduce), evolution will also keep the alignment between every single organ and the function characterising it (since malfunctioning of the organ leads to malfunctioning of the organism).

The very concept of function, which, according to some would necessarily refer to the intention of a human being (a conscious mind, who creates an object for a certain purpose, or assign a purpose to an existing object choosing to use it for a certain purpose (see, for example, Searle [Sea95], 13), can also be generalised in such a way to be independent from such a reference<sup>3</sup>. Besides being applicable to artifacts and biological organisms, the design stance can also be applied to social organisation, both in its purpose-based version

---

<sup>3</sup>For example, Nozick [Noz93] defines the notion of a *n* on the basis of the concept of an homeostatic system, which he defines as follows: : “[An homeostatic system] maintains the value of one of its state variables *V* within a certain range in a certain environment, so that when *V* is caused to deviate some distance (but not an arbitrary long distance) outside that range, the values of the other variables compensate for this, being modified so as to bring *V* back within the specified range.”

Consider for example how an increase of bodily temperature may lead to sweating, which lowers the temperature, or consider how an increase in the temperature of a house may start air conditioning, which goes on until the temperature has fallen to the established level

Nozick defines then the notion of a function as follows: “*Z* is a function of *X*, when *Z* is a consequence (effect, result, property) of *X* and *X*'s producing *Z* is itself the goal state of some homeostatic mechanism *M* . . . , and *X* was produced or is maintained by this homeostatic mechanism *M* (through its pursuit of the goal: *X*'s producing *X*)”. According to this definition we may say, for example that the function of thermostats is that of keeping temperature within the specified range, since stabilising temperature is the result which is obtained through the process of designing and building thermostats, a process which tends to make so that thermostats are built which can stabilise temperature (thermostats designers constantly endeavour to improve the performance of the thermostats they design). In the same way, the function of lungs is that of absorbing oxygen, since this is the result natural evolution, which tends to make so that lungs are able to absorb oxygen (through survival (and therefore reproduction) of the individuals whose lungs can absorb oxygen).

and in its evolutionary one. In fact, the behaviour of a public or private organisation can be explained on the basis of the functions performed by it (or by its components). In their turn, such function can be originated or by the original design of the organisation, or by the way in which organisations of this type survive and reproduce.

So, I will explain that a commercial company (usually) produces profits by providing the following reasons: (a) the company has been created by its partners with the purpose of producing profits, and (d) the evolutionary mechanisms of a market economy eliminate companies that do not produce profits and lead to the imitation of most profitable companies.

Finally, we need to observe that the design stance, through allowing us to make explanations which are normally correct (usually an artifact realises the purposes of its designer, and usually a biological organism works in such a way as to promote its persistence and reproduction) may be fallible in particular cases. In the case of intentional design, the designer may have made mistakes; in the case of biological organisms, reproduction mechanisms may have produced counterproductive variations (which weaken the new organism); in the case of an organisational structure, both types of degeneration may take place. Moreover, even when an entity would work appropriately in its original environment, it may malfunction with regard to a modified environment.

## 4 The intentional stance

Let us now move to the third and most controversial perspective, that is the intentional stance. First we need to specify that here the term “intentional” is used in the technical meaning it has in philosophical language, where it typically refers to the relationship (also called “aboutness”) between mental states or linguistic objects, and the things to which they refer to. We are not restricting the meaning of “intentional” to being deliberate, or on purpose. Thus, also beliefs, desires, hopes and fears are intentional states, since they refer to what is believed desired, hoped or feared (on intentionality, cf. Dennett and Haugeland [DH87]).

When looking at an entity from the intentional stance, we are explaining the behaviour of that entity assuming that it has certain cognitive states. Within cognitive states we include both epistemic states (information on how things are) and conative states (information on what to do). Typically,

we assume that the entity we are examining is trying to achieve certain objectives (goals) or to apply certain instructions (intentions) on the basis of certain representations of its environment (beliefs). The behaviour of an intentional entity will then be explained as resulting from that entity applying the instructions it has adopted (according to its intentions and its beliefs on the existence of the relevant circumstances) or from its attempting to achieve its goals (through means it believes to be appropriate).

For example, to understand and forecast the behaviour of a chess-playing computer system, usually one can apply neither the design stance (consider the functions played by the modules of the system and the programming instruction they contain) nor the physical stance (consider the electrical state of the components of the computer). In fact, the adversary of such a system knows nothing of its software structure and even less of the electrical status of its hardware components. Moreover, even a programmer involved in building the system cannot anticipate its way of functioning by mentally executing the programming instructions it includes (the system is too complex for one to be able to do that). One can only predict the behaviour of the chess-playing system by attributing to it goals (winning the match, attacking a certain piece, getting to a certain position), information (about what moves are available to its adversary), and assuming that it can devise rational ways to achieve those goals according to the information it has. According to Dennett, the intentional stance works as follows:

first you decide to treat the object whose behaviour is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. (Dennett [Den89], 17)

We may adopt the intentional stance first of all with regard to human beings. It represents indeed the usual way in which we understand and forecast other fellows' behaviour.

The explicability/foreseeability of other people's behaviour does not oppose recognising that people have a spiritual life (they have their own purposes, beliefs, desires, intentions), but on the contrary is based upon this recognition. This is what allows us to have beliefs like the following: the electrician has turned off the switch before starting to work since he wants

to avoid the risk of being electrocuted; the broker will buy certain shares since she wants to make a profit, and she foresees that those shares will increase in value; the attorney will provide a certain argument, since he believes that this will lead the judge to decide in favour of his client; a party will accept to buy a merchandise at a high price since she needs it and believes that nobody else can provide it.

We may however adopt the intentional stance also with regard to animals. So, we may assume the following: the ape moves the chair below the cask of bananas since it wants to reach the bananas; the dog cuts across the hare's path since it wants to catch it; the bird collects a straw in order to use it in building its nest; the fly flies off since it has seen the hand above it and wants to escape; the ant is dragging a crumb since it knows that it is food and it wants to bring it to its nest.

In some cases, the intentional stance may also be appropriate towards vegetables. For example, an appropriate answer to the question why a plant has started producing certain toxins may be that the plant knows that a parasite is attacking it and it wants to defend itself. Similarly, an appropriate answer to the question of why a virus has changed the chemical structure of its protein cover may be that the virus tried to resist to the antibodies of its host organism.

As these examples show, adopting the intentional stance toward an organism does not exclude that, when we have appropriate theoretical models, and the time and the energies to apply them, we may study the same entity by using the design stance or the physical stance. For example, we may keep the idea that the virus is trying to resist to the antibodies, and at the same time, adopt the design stance to distinguish the functions which are performed by the protein cover of a virus (protecting its DNA against external chemicals), and adopt the physical stance to study the chemical or physical interactions between the protein cover and antibodies. Therefore the lower stances are not alternative to the intentional stance, but rather complete it, correct and specify the abstract and synthetic results which are provided by intentional interpretations.

Obviously, the intentional stance is the more useful the more the concerned entity is able to select, in a large range of possibilities, the behaviour that it is most appropriate to achieve its objective (to realise its function). The intentional stance, on the other hand, is neutral with regard to the process through which the concerned entity adopts its objectives and chooses how to pursue them, in the existing circumstance. This process may consist

in explicit and conscious planning or it may also consist in feedback driven by reinforcement. In the latter case, the entity, given certain environmental stimuli experiments different reactions, and tends to repeat those reactions with activate the reinforcer (pleasure or any other state which is somehow related to the achievement of the functions of the entity). Finally the selection process may also consist in evolution: the entity reproduces itself, or reproduces certain of its components, or certain of its behavioural reactions, which mutation, and the mutants are preserved which are more conducive to the functions of the entity.

The intentional stance is adopted wrongly only when the concerned entity lacks the ability to make determinations which are appropriate to its objectives, in the context in which it is pursuing such objectives. For example, it may be wrong to adopt the intentional stance toward the forces of nature, such as the sea or the wind, while it is correct to apply it towards living micro-organisms, at least as a first approximation.

The intentional stance is certainly appropriate in regard to complex artifacts, and in particular, in regard to computer systems.

To use the classical example by Dennett, let us consider again computer systems for chess playing. These softwares can play at different levels of competence, and some can compete with chess masters. A few years ago one such systems, called Deep Blue, developed by IBM, achieved fame since it defeated Kasparov, the world champion. The victory of Deep Blue started many debates on the connection between artificial and human intelligence, and on the chance that in the future humans may be overcome by machines. Here we will consider a different question, that is the attitude that a human should adopt when interacting with such a system.

Imagine that I am a chess champion and that I face Deep Blue trying to avenge Kasparov's defeat. In chess, every one of my moves depends on my expectations concerning the moves of my adversary. These expectations are largely based, when I am facing a human, on attributing certain intentions, strategies, objectives, beliefs to my adversary. For example, I may assume that my adversary, to win the match, intends to eliminate pieces of mine, when she may do that in such a way that her losses are inferior to mine (and the operation has no negative side effects). I may therefore foresee that when my adversary believes that a certain strategy allows her to achieve this result, she will implement this strategy. Finally, when my adversary has a considerable competence in chess playing, I may forecast that, if I give her such an opportunity, she will seize it, to my loss. Therefore, I will try to

avoid giving her such an opportunity.

How shall I reason when I face a computer system, rather than a human being? Shall I adopt the same strategy for interpretation and prediction (the intentional stance), and therefore attribute intentions and beliefs to Deep Blue, or shall I adopt a different point of view?

It is completely impossible for me to adopt the physical stance: I should know the precise electrical conditions of every component of the computer on which the program is running. It would be like trying to foresee the behaviour of my human adversary on the basis of the physical and chemical condition of every cell of her brain.

Also the design stance will not take me very far. I cannot go beyond the generic hypothesis that the designers of Deep Blue wanted to make a program which was good in playing chess. Since I do not know the internal structure of the software it is very difficult for me to conjecture through what programming architecture the designer has intended to achieve this result, and in any case it would be impossible for me to make this conjecture so precise that I can use it to explain and foresee the behaviour of Deep Blue. It would be like me to try to interpret the behaviour of a human adversary on the basis of the function of the human brain and of its components.

In conclusion, the only perspective from which I may try to interpret/foresee the behaviour of Deep Blue is the same that I can adopt with regard to a human adversary: attributing intentional states to it and adopting a strategy which may be winning over the strategy that I assume Deep Blue is following.

We need to consider the issue of intentionality not only with regard to software systems, but also with regard to physical robots, i.e., intelligent systems made of hardware e software, which operate to some extent in the industry, but which are likely to enter soon our houses as toys, cleaners, servants, etc. How shall we interact with automata which can execute their functions with autonomy and intelligence, process visual and sound stimuli, interact linguistically? It seems that our only chance is that of attributing them mental attitudes (beliefs, intentions and possibly even emotions) and interpreting correspondingly their behaviour.

Obviously, the intentional stance may also be adopted with regard to public and private organisations: Microsoft, IBM, the Italian State, the European Union, etc. This perspective will be appropriate to the extent that the concerned organisation is able to act for achieving its purposes, by choosing means that are appropriate, on the basis of the information accessible to it. So, I can explain/foresee the behaviour of Microsoft (for example the de-

cision to market a new version of its operating system, through a new legal and commercial model), by attributing Microsoft intentional states. I will assume that Microsoft intends to achieve certain results (facilitating updates, crashing competitors, binding users, reducing piracy), since it has certain expectations concerning economic and technological trends, and the behaviour of other actors (consumers, competitors, etc.). I can adopt this stance even without knowing what individuals, within Microsoft, will have those intentions and expectations, and even if I believe that such attitudes cannot be fully attributed to any individuals.

Finally, the entity which I consider from the intentional stance can be a mixed subject, that is a combination of human, electronic, and organisational components. Consider for example an e-business structure which includes different components: a software which interacts with customers (drafts and sends sale offers, receives and confirms acceptance by the customers, controls and monitors the execution of the contract, accepts and processes some types of complaints), programmers write and modify the software, employees parametrise the software (for example, they select what items to sell, establish and change descriptions and prices for sold items), managers define objectives and tasks for programmers and employees (on the basis of aggregated data).

No component of the e-business organisation has a precise view of all information processed by that organisation: managers do not know neither the prices of items nor the instructions in the programs, programmers and employees do not know the strategy of the organisation nor they know the what specific contracts with what customers are made by the software, the software does not have the information on the basis of which its parameters are modified. However, the organisation as a whole appears to function rationally: it pursues its objectives (selling certain types of products, while maintaining its market share and making profits) keeping into account all information available to it (ranging from the general trends of customers' taste to the address of one individual customer). The system as a whole appears to be a unit of agency (for example, in regard to people accessing its web site) and to such unit we may attribute certain intentional states (for example, the "will" to make a certain contract and the knowledge of its contents and presuppositions), even when such intentional states cannot be attributed to any element considered in isolation.

In particular, the intentional stance is the only perspective from which one usually may hope to understand and forecast what a SA (software agent)

will do. This forecast cannot be based upon the analysis of the computational mechanism that constitutes the SA, and on the pre-determination of the reactions of this mechanism to all possible inputs. The user of a SA will normally have little knowledge of those mechanisms, and even the programmer who built the SA will be incapable of viewing the SA's present and future behaviour as the execution of the computations processes which constitute the SA. The overall interpretation of the SA's behaviour will be based upon the hypothesis that the SA is operating "rationally", by adopting determinations which are appropriate to the purposes which have been assigned to it, on the basis of the information which is available to it, in the context in which it is going to operate, that is according to the intentional stance.

This assumption of rationality needs not to be absolute. On the contrary, it may be integrated by the knowledge of the limitations of the capabilities of the SA (so that one may also explain why the SA fails to behave rationally under certain particular circumstances). This explanatory model is similar to the strategy we adopt in regard to humans: we can interpret and forecast other people's behaviour by combining the general hypothesis of their rationality with the knowledge of the limitations and idiosyncrasies of each individual. If the intentional stance needs to be adopted by the user of a SA, *a fortiori* it can be adopted by the SA's counterparties in exchanges (where the SA is acting on behalf of its user/owner). The counterparty of the SA cannot even try to understand the behaviour of the SA by analysing its software code (the code is usually inaccessible, and in any case it is too complex to be studied in time), nor by wondering what intentions of its user, codified in this software, the SA may be expressing.

Consider for example, an animated shop assistant, who appears as a three-dimensional cartoon endowed with body language (face expression, gestures, etc.) and speech, which leads a client into a virtual shop (of antiques, used cars, etc.), presenting him the products, questioning him about his needs, suggesting certain choices, and proposing certain contractual terms. Consider also the case of a virtual tour-operator, possibly speaking through the user's mobile phone, asking her about her need, and proposing her to buy certain tickets, on certain conditions. Finally, consider a SA operating in a dynamic market environment, and contacting both people and other SAs in order to find the best deals. For the interlocutors to such SAs, the only key to understanding the behaviour of the latter will be the hypothesis that SAs, in order to achieve the objectives assigned to them, and by using the knowledge they have, will get to the determinations which they declare to

their counterparties, according to existing linguistic and social conventions. So, the assumption of rationality (relative to the cognitive states of the SA) needs to be integrated with the knowledge of the limits of such rationality, still provides the default background for understanding the SA's behaviour.

## 5 The nature of intentional states

Before concluding our discussion of the intentional stance, let us approach the difficult issue of determining the ontological status of intentional attitudes: what is the reality which is described by asserts which attribute an intentional attitude (asserts affirming that an entity desires, believes in, or intends to do something). What inferences can we draw from such asserts, under what conditions are they true or false? This is a very difficult philosophical issue, with regard to which we can only make some short and provisional considerations.

The idea of the intentional stance, as resulting from the Dennett's quotation above, seems to lead us toward the conclusion that cognitive states only exist in the eye of the observer. They appear to consist in a particular way of looking at an entity, which cannot be translated into internal features of that entity. To say that an entity has certain cognitive states (goals, information, beliefs, desires, etc) would just mean to affirm that its behaviour is explicable and predictable according to the intentional stance, that is by attributing cognitive states to it (and postulating that the entity can behave rationally on the basis of those cognitive states). This leads us to a kind of behaviouristic approach to intentionality: it is the behaviour of a system which verifies or falsifies any assertions concerning its intentional states, regardless of its internal conditions. So, for it to be the case that my chess-playing system "wants" to eat my tower, it seems sufficient that by attributing this goal to the system (and assuming that it can act in such a way as to achieve its goals) I can foresee its behaviour (anticipate future moves). In the same way, we may say, for it to be true that an amoeba "wants" to ingest some nutritional substances, it is sufficient the ascription of this will allows me to explain effectively the behaviour of the amoeba (the fact that it moves, approaching where such substances are present, and then absorbs them).

From this perspective, if two entities behave exactly in the same way in every possible situation (and their behaviour is therefore explicable on the basis of the same ascriptions) we must attribute them the same intentional

states, even if their internal functioning is completely different. Let us assume, for example, that two programs for playing draughts work exactly in the same way (they make the exactly the same moves in the same conditions), but that they work on the basis of different principles. The first chooses its moves on the basis of a calculation of the chance that they contribute to achieving a favourable situation, considering possible replies of the adversary. The second consists of a huge table, that connects every possible situation in the board to a specific move. From the point of view we have just considered, since the two systems behave exactly in the same way, it seems that we cannot attribute intentional states to the first and deny them to the second.

However a different approach is also possible. One may take a realistic view, which asserts that cognitive states concern to specific internal features of the entity to which they are attributed. So, to establish whether an entity truly possesses cognitive states (goals, beliefs, intentions, etc.) one needs to consider whether there are internal conditions of that entity that represent epistemic states (beliefs) and conative states (desires, goals, intentions), and whether there are ways for that entity to function which implement rational ways of processing epistemic and conative information. The behaviour of the concerned entity would only be relevant (to the possession of intentional states) as a clue to its internal functioning.

Obviously, a realistic attribution of intentional states to hardware or software artifacts or to organisational structures, assumes that we can identify appropriate internal states of such entities.

Following this line of thinking, we can say that an internal state of a certain entity (for example, the presence of a certain chemical substance in the circulatory system of that entity, or the presence of certain character strings in a certain variable or data buffer) represents an epistemic state, and more precisely, the belief in the existence of certain situations, when the concerned entity:

- adopts that state on the basis of those situations (in such situation the entity's sensors are activated and this starts a causal process that leads the entity to adopt the state) and
- having that state contributes to making so that the entity behaves as those situations require.

Therefore, the internal state of an entity is a belief concerning the existence

of certain external situations (or, if you prefer, it represents or indicates such situations to the entity having that belief), when

- there is a covariance between the internal state and those situations, and
- this covariance enables the entity to react appropriately to the presence of those situations.

We cannot approach here this difficult issue (for the idea of covariation, cf. Dretske ([Dre86]), and for a discussion of the literature, cf. Davies [Dav98a], 287 ff). Let us just observe that we may look from this perspective to computer systems, and in particular, to SAs. Ascribing epistemic states to computer systems would allow us, at least in some cases, to find an appropriate legal discipline without new legislation, and moreover to distinguish clearly those situations when possessing, or causing others to possess, epistemic states is relevant to the law.

Consider for example the action of inserting a name in the buyer's slot, in the process of ordering something from a web site. This registration certainly tends to covariate with the name of the person who is making the order, and it enables the site to behave in a way that is appropriate to a contractor (and not to a person who is impersonating somebody else, without the consent of the latter). Therefore, we may say that the site "believes" that the registered name is the name of the person who made the registration (or of a person that authorised the latter). When the name provided is different from one's real name, we can say that the site has been deceived, i.e., that it has been induced into having a false representation of reality. This would allow us to apply, at least analogically, those legal rules concerned with deception, also to interactions with computer systems (a step that in most countries jurists were reluctant to do, thus requiring a specific legislation on computer fraud).

The idea that a computer system can have cognitive states may also be extended to conative states. One may say that an entity has the goal of realising a certain result (in more anthropomorphic terms, that it has the desire), when there is an internal state of the entity such that:

- while the entity has that internal state it will tend to achieve that result, and
- when the result is achieved the internal state will be abandoned (or modified so that it stops producing the above behaviour).

For example, when a biological organism (even lacking a brain, or having a rudimentary one) lacks the substances it needs for surviving, it comes to have certain biochemical states that push it to search for and ingest food, and those states cease to obtain when the lacking substances have been reintegrated. We may therefore say that such biochemical states represent to the organism the objective of obtaining an adequate nourishment. Similarly, assume that a SA has been given the description of certain goods and that, on the basis of such description, it starts operating in order to buy those goods, until the goods are purchased (which will lead it to remove the description from its task queue). Under such conditions, we can certainly affirm that the SA has the objective, goal, or end of buying these goods.

We may also say that a system wanted to perform a certain action, if there was an instruction in the system, which prescribed the system to perform that action. Finally, we may say that a system has “wanted” to adopt certain behaviour, if that behaviour resulted from an internal process, intended to make so that the system achieved its goals, on the basis of its epistemic states.

For example, if a SA has adopted the goal of damaging somebody or something (for example, the goal of making a system crash), and has chosen to perform an action producing that result, as a way of producing that result, we may say not only that the SA wanted to take that action, but also that it wanted to produce the result, i.e. that the damage was deliberately caused by the SA.

Let us conclude this discussion of cognitive states in computer systems by affirming that, from a legal perspective, there is no need to make a choice between the two views we have just considered, that is between:

- viewing cognitive states as mere interpretations of the behaviour of a system, and
- viewing cognitive states as internal conditions of the system.

The two views are, first of all, linked by a causal connection: usually an entity can behave in a way that corresponds to a certain cognitive interpretation, exactly because it has internal states of the type we have described. Moreover, from a legal perspective, the two conceptions are complementary. On the one hand, the idea of cognitive states as interpretations of an entity’s behaviour focuses on the attitudes of external observers (what beliefs, goals and intentions do counterparties attribute to the entity?). On the other hand

the idea of cognitive states as internal states refer to the point of view of the entity itself, or of those who can inspect its internal functioning (does the entity really believe what the counterparty assumes it believes, and has it has the goals which the counterparty assumes it has?).

Some authors have linked the notion of intentionality to the notion of consciousness or awareness. For example Searle ([Sea89], 208) affirms that “Roughly speaking, all genuinely mental activity is either conscious or potentially so. All the other activities of the brain are non-mental, physiological processes”. According to this author “The ascription of an unconscious intentional phenomenon to a system implies that the phenomenon is in principle accessible to consciousness (Searle Searle1990CE, 586). Searle’s connection between intentionality and consciousness forces us either to renounce to the intentional stance in regard to all non-human entities (unless we want to trivialise the notion of consciousness). We need to reject such connection to approach a (real and virtual) environment where we must increasingly face autonomous (biological, organisational, electronic) systems.

Possibly, we may rephrase the problem of the connection between intentionality and consciousness as concerning the distinction between direct and reflexive intentionality. The first, as we have seen, consists in the fact that the behaviour of an agent is explicable/foreseeable through the ascription of intentional states. The second consists in the fact that the concerned entity can look at itself from the intentional stance and view itself as the bearer of beliefs, goals, intentions, projects, and to make its behaviour approximate this ideal (cf. Dennett [Den97], 119 ss.)

Such a capacity can be fully attributed, besides than to humans, also to such organisational structures which can critically examine their own policies, choices, ways of functioning. On the other hand the intentional stance remains applicable also to entities (such as animals and SAs), to which it seems that we cannot attribute reflexivity.

In the same way, it seems that we need to reject a necessary connection between intentionality and normativity, i.e., the idea that attributing intentionality to an entity presupposes that such entity has the capacity of following norms or rules, even when they clash against its desires. Even if we do not consider the difficulty of distinguishing the cases of real normativity from the conflicts between impulses (for example between altruistic and egoistic impulses) which may also occur in animals or artificial systems, we need to consider that the intentional perspective legitimately be adopted (and is frequently adopted) also towards entities that certainly experience no “sense

of duty” similar to that which is experienced by most humans (at least in some occasions).

The consideration we have developed above lead us to positive conclusions concerning the possibility of attributing to artificial entities the specific type of intentionality which seems to underlie the execution of speech acts, and in particular declarations (of will or intention). From an intentional perspective, this consists in explaining a certain behaviour (usually the performance of an utterance), according to the view that the agent intends to produce certain normative results through the utterance of a certain statement, believing that this utterance (accompanied by such intention) will produce such results. There are no reasons for not viewing from this perspective, for example, the case of a SA stating that that it wants to purchase a certain item (see the Lovely Rita example).

## **6 The intentional stance and the law**

In the previous pages we have observed that when we are interacting with complex entities we need to go beyond the physical stance: we need to adopt also the design stance and the intentional stance. The possibility of adopting such stances is a fundamental condition of social life. If we could not look at the world also from the design stance, we would not be in the condition of surviving: we could only eat the foods which have undergone a full chemical analysis, we could only use objects of which we know perfectly the internal structure, etc. If the intentional stance was precluded, it would be impossible to participate in social life. We could build expectations concerning the actions of peoples and animals only on the basis of a complete scan of their brain, or at least of a full functional map of the brain’s components.

Similarly, if we had only access to the physical stance, we could interact with a corporate body only of the basis of a complete knowledge of its organisation chart and of all its, formal and informal, circuits of communication and power. We could interact with an electronic system only on the basis of a full knowledge of all of its software and hardware components. We would have the same limitation with regard to mixed organisations, where information processing is allocated partly to electronic devices and partly to humans.

The law is not neutral concerning the need to allow and even to promote the adoption of the design and intentional stances. On the contrary, it

frequently intervenes to support and guarantee reasonable any expectations (reliance) that people have as a consequence of adopting such stances (Jones [Jon00], Castelfranchi and Falcone [CF03]).

Let us first consider the design stance. When we look at an artifact that seems to embody a certain design, why do we expect that it works according to that design? Why are we ready to take the risk that the behaviour of the artifact does not correspond to such design?

Different factors converge in supporting one's expectation that the behaviour of the object corresponds to what is expected according to its assumed design, but among those factor there are also some legal rules. For example, the law ensures some protection to the expectations based upon the design stance in the following cases: it requires the seller to guarantee that the thing has no faults; it states the producer is liable for damages caused by malfunctioning; it requires that the owner or guardian of a thing is liable for damages caused anomalous behaviour of that thing.,

As a consequence of the rules we have just mentioned (and of may other legal rules), a person expecting that an artifact behaves according to its apparent design is protected when such expectations are disappointed. We need to observe, additionally, that this protection is realised through putting the obligation to compensate damages upon the subject who could prevent such disappointment. This normative guarantee leads to a factual guarantee to the extent that it induces the obliged person to behave in the way that corresponds to other people's expectations. The combination of the two aspects we have just indicated (on the background provided by non-legal mechanisms: reputation, social conventions, etc.) makes so that we can enjoy a certain degree of trust in the artifacts with which we need to interact.

Let us now consider the expectations that we form when looking at social reality from the intentional perspective. Why should we interpret other people's expectations by ascribing beliefs, desires and intentions? Why do we take the risk that interpretation and forecasts based upon the intentional stance are disappointed, that the person to which we attribute a certain mental state does not behave according to the expectations that are grounded upon such ascription.

First of all, we need to consider that psychological components (intentional states) play an important function in many legally relevant acts, from crimes, to torts, to contracts and other juristic acts. The recognition of such components, has two aspects. On the one hand the legal effects of an act may be conditioned to the presence of the following psychological states

(usually intentions and beliefs): the intention to make a certain declaration, the intention to realise the normative states one declares, the (wrong) beliefs concerning the reasons which determined such intentions. On the other hand, the effects of one's acts are also conditioned to the intentional state that the counterparty ascribes to the author of the act, of the basis of the available clues: first of all the content of express declarations, but also other behaviour of its author. If there is a difference between the mental states which are possessed by the agent and the mental states which are attributed to it by the counterparty, the decisive criterion can be represented by social conventions, which define the meaning that one party can legitimately ascribe to the behaviour of the counterparty.

Such triple intentional qualification of one's behaviour (the point of view of the author, of the counterparty, of the existing social conventions) leads to the conflicts which are well known to the students of the the formation of contracts and of the protection of reliance. We cannot address such issues here, but we need to observe that the law seems to be incompatible with those theories which are completely sceptic in regard to intentional notions, such as the so called eliminative materialism or behaviourism. The law, on the contrary, is attentive to the issue of the protection of reliance (trust): one party's erroneous belief that the counterparty made a certain declaration (which the counterparty did not make) or had a certain intention (which the counterparty did not have) can sometimes produce the same effect that the counterparty's making that declaration or having that intentions would produce. This happens to protect the party who was justified in believing that the counterparty made that declaration or had that intention , and this is the case when this belief corresponds to existing social conventions.

This aspect of the legal discipline of contracts has sometimes been considered as the symptom of the passage from a subjective to an objective perspective in evaluating legal acts, as the abandonment of psychology in favour of sociology. We believe that the protection of reliance does not represent a rejection of the intentional or psychological aspects of human actions, but is rather the attempt to facilitate and secure the possibility of giving an intentional interpretation to the actions of other people. I can rely upon my attribution of certain intentional states to my partner (her will to sell a certain good, her intention to perform a certain task, her belief in what she affirms), on the basis of a correct interpretation of the clues which are provided by my partner, according to the existing social conventions, since I know that, even if my ascription was wrong, it would still determine the

legal effects which would take place if it were true.

Moreover, I know that my partner knows the legal evaluation of her own contractual behaviour (and in particular, knows how the law protects my reliance). Consequently, I may expect that she will adopt all care which is needed to prevent possible misunderstandings, and I therefore can assume that she will really have the mental states that she appears to have.

Also other legal rules tend to ensure the agent's capacity to attribute intentional states to his partners in various social interactions. Consider for example legal rules which punish the malicious communication of false information: here the law protects the reliance in other people's assertions, transferring the damages that one has suffered for relying on false assertions upon the author of such assertions.

Frequently the law also considers higher level cognitive states. For example, one contractor's intentions need to coincide with his beliefs concerning the other party's intentions. Therefore, when I am concluding a contract, I cannot, in good faith, attribute one meaning to one clause and at the same time believe that my counterparty attributes a different meaning to the same clause. Similarly, in mistake and deceit, the law attributes relevance to one party's knowledge that the counterparty does not know a certain fact (if the contractor knows that the counterparty is mistaken in regard to an essential term of the contract the contract may be voidable).

In conclusion, according to the model we have developed, the attribution of intentional states to an entity if grounded upon the hypothesis that the entity possesses the capacity to act rationally (efficiently) in pursuing its objective, on the basis of the information that is accessible to it. The statements that attribute to an entity a certain intentional attitude are true when this entity really possesses an intentional state that performs the function that is proper to the intentional attitude which is attributed to that entity.

The same assertions are correct or justified when the behaviour of the entity provides sufficient clues to conclude that the entity possesses the attitude which is attributed to it. Whenever an entity appears to possess the capacity to process information which is the precondition of its having intentional states, we are justified in looking to it from the intentional stance, and consequently we are justified in attributing to it aims, intentions and beliefs.

As we observed, adopting the intentional stance is not an arbitrary choice: such perspective represents usually the only possible viewpoint to explain and foresee the behaviour of complex entities which can act teleologically. Consequently, the law should not refuse to acknowledge the intentionality of

artifacts, it should rather give it legal relevance whenever this may be useful. In particular, this approach would allow us to approach better automatic contracts, and in general, all interactions between humans and computer systems or between such systems. In fact, it would enable us to apply to computer systems the same set of well known intentional legal notions (will, reliance, mistake, deceit, malice, etc.) which is currently used for humans, and so to provide more flexible, intuitive and tested legal solutions than those which may be provided by new legal notions, purely behaviouristic, specifically intended for computer systems.

The recognition of the intentionality of computer systems (and of mixed systems, including both humans and computers) would imply two sets of consequences:

On the one hand who is interacting with such a system would be authorised to attribute to the systems the intentional states that it appear to have, and in particular, those intentional states that the systems declares to possess or that are presupposed by the speech acts it has accomplished. According to the principle of protection of reliance, the owner of the system will not be able to avoid that the systems is assumed to have those intentional states that (a) have been really been attributed, in good faith, to the system by the counterparty and (b) are reasonable interpretation of the behaviour of the system, according to the conventions which are applicable to the concerned interaction. For example, if a SA performs a speech act which appears to be a statement of fact, I will assume that the SA believes what it is declaring, and I may consider to have been deceived if the SA choose to provide me with false information (and accuse the SA of lying, with the consequences this implies against the owner of the SA, for example in regard to tort liability). Similarly, if a SA performs a declaration of will or intention (typically, a contractual offer or a declaration of acceptance) I may assume that the SA intends what it declares, and the owner of the SA will not be able to avoid the effects of the action of the SA by affirming that he had not the intention of performing that action.

On the other hand, the counterparty of a computer system will not be able to avoid those interpretations of the behaviour of the system which: (a) correspond to intentional states really possessed by the system (b) are attributable to the system on the basis of conventions which are applicable to the interaction at hand. For example, I will not be able to avoid the attribution of certain contents to the contract I have made with a SA when the SA really had the intention of making such a contract and moreover the

same contents may be attributed to the contract according to the existing conventions. In the same way, I will not be able to avoid the usual consequence of the fact that the SA made a decisive mistake (the voidability of the contract), when I should have been aware of that mistake on the basis of the behaviour of the system. As this last example shows, the intentional stance can be adopted not only towards the intentions of a computer systems which are directly expressed in the declarations of the systems, but also towards the mental states that are presupposed by such declarations. So, when a SA make a contractual declaration (for example, it declares the willingness to by a certain good for a certain price) I am authorised not only to attribute the SA the intention of making such declaration, but also the mental states that are presupposed by such intention (for example, the belief that the object of the contract has all features which have been advertised or which are normally possessed by objects of that type).

To conclude with a broad statement, we may affirm that in a world that is characterised more and more by man-machine interactions, we have to face the alternative between on the one hand objectifying also human relations, submitting them to a purely behaviouristic legal discipline, and on the other hand spiritualising also relation with or between machines, viewing them from the intentional stance. It seems to us that the second approach may be the one that best contributes to realise a social environment (even in the virtual dimension) where people can feel at ease and explicate their social abilities, also in legal relationships.

Finally, let us remark that, though here we are focusing on computer system, the intentional stance also is very important is also in approaching collective organisations. Towards such organisations the intentional stance can be adopted by both outsiders and insiders: to both the organisation appears to be a subject having its own objectives, and having processes for acquiring and processing information (theoretical rationality) and for using it in decisions (practical rationality). Therefore, we reject both the theories that view the subjectivity of organisations as a mere fiction and the theories that view it as resulting from a purely legal mechanisms. The recognition of the autonomy of organisations (as subjects which are distinguished from their members) seems to be based upon the phenomenon which we tried to show in the previous pages: the need to adopt the intentional stance to explain and predict the behaviour of entities which are able to know, choose and act. Moreover, attributing an intentional state to an organisation does not presuppose necessarily that the same state can be attributed to a specific

individual member of the organisation.

## 7 The delegation of cognitive tasks

Our discussion of the issue of intentionality and cognition in computer systems has led us to the conclusion that SAs (as other computer systems) can be attributed intentional states, and can perform cognitive processes. Viewing them from this perspective, is the only effective way of approaching SAs for most purposes. Therefore, delegating them cognitive tasks is to be the appropriate way of using them, and viewing them as having been delegated such tasks is the appropriate way of interacting with them.

From this perspective the reason why the effects of what is done by a SA will fall upon the user is not the fact (or the fiction) that the user has wanted or has foreseen the behaviour of his agent (cf. Sartor [Sar02]), but rather the fact that the user has chosen to use the agent as a cognitive tool, and has committed himself to accepting the results of the cognitive activity of the agent. So, the legal efficacy of the action of the SA (especially in the contractual domain) will exclusively depend upon the will of the user (as it can be reasonably construed by the SA's counterparties), but this is only the will to delegate certain cognitive tasks to the SA, not the will of performing of all specific actions that will be performed by the SA on behalf of its user. Since the user intends to rely on the SA's cognition, and this is known to potential counterparties, the fact that the user is responsible (in the sense that he will bear the rights and duties resulting from the activity of the SA) does not exclude, but rather presupposes, the legal relevance of cognitive states and processes of the SA. So, the fact that SAs have their own cognitive states and perform cognitive processes which are not attributable to the user (and the awareness of that, by users and other parties) distinguishes SAs from other objects or tools, also from a legal perspective.

It is true that the phenomenon we are discussing (cognitive delegation to an artifact) is not completely new: this happens (though in a trivial way) whenever one is using a calculator or a computer system as a decision aid, before making a contract or taking any legally relevant decision (consider a shopkeeper using a computer to track sales, compute prices and taxes, or to check somebody's credit card). However, usually this only concerns the preliminary steps of a deliberation, and leads to cognitive results that will be appropriated by the person who deliberates (e.g. who concludes a contract).

So, usually there is no need for the law to give a separate consideration to automatic cognition. When, for example, a software mistake determines a mistake of the user, it is sufficient that the law takes the latter into account. This is not the case, however, when a SA is charged with accomplishing a legal activity directly, i.e., when “no natural person reviewed each of the individual actions carried out by such systems or the resulting agreement.”, as it is said in art. 12 of the Unicitral Document “Legal aspects of electronic commerce. Electronic contracting: provisions for a draft convention”<sup>4</sup>. Under such circumstances we need to address directly the issue of cognition performed by artifacts, and consider what its legal relevance may be, according to the nature of those artifacts, and the (reasonable) expectations of their users and interlocutors.

One may argue that even the issue of artificial cognition is not completely new to the law: there has been a progressive development of machines used for performing cognitive tasks finalised to legal results, even without a user’s review: from automated vending machines, to cash dispensers, to EDI contracting, to computer contracting through Internet sites (the classical reference for Italian legal doctrine is Cicu [Cic01]). We are indeed happy to take on board this observation, but as an invitation to rethink the legal discipline of all cognitive tools, and to find a conceptual framework that, though more needed for more complex cognitive tools, such as SAs, will also apply to simpler automata, like the ones we just mentioned.

In the following we will consider some legal implication of the approach we have just sketched, in regard to two different areas of the law, tortious liability and contracts (for a review of other legal issues related to SAs, see Cevenini et al [CGV02]).

## 8 SAs and tortious liability

Some issues concerning liability for damages produced by SAs are common to other technological objects. Let us consider, first of all, the issue of identifying a custodian of a SA, in order to make him responsible for damages caused by the SA. With regard to material things this is often easy: this is their owner, unless the latter has transferred control over the thing to somebody else (e.g. a borrower or a lessee). However, in regard to SAs, we must

---

<sup>4</sup>United Nations Commission on International Trade Law. Working Group IV (Electronic Commerce). Thirty-ninth session. New York, 11-15 March 2002).

consider that different subjects may “own” different aspects of a SA. If the SA implements a patented technology, then the patent holder owns this. If the SA includes (as usually is the case) copyrighted software, then this (the software in itself) is “owned” by the copyright holder. If the SA includes a database of information, then this is “owned” by the collector and organiser of the data. If the SA contains personal data (e.g. data concerning its user), then one may possibly argue that such data is “owned” by the data subject, according to data-protection law (this would be the proper approach at least when one adopts a proprietary approach to personal data, as suggested for example, by Lessig [Les99], 142 ff). Finally, if the SA is under the control of some user, then the user may be said to “own” the particular combination of technology, software and data, which constitutes the SA.

Note that we have always put the expression “owned” between inverted commas, to indicate that each one of those entitlement should not be construed as the usual property right over material things, but as denoting a peculiar cluster of rights powers and duties, as established by the law of intellectual property and data protection, or by contractual relationships (on ownership of SAs, from a computer science perspective, cf. Pitt et al [PMC01], Yip and Cunningham [YC02]). In any case, since the user has no control over certain aspects of his SA, it may be unfair to regard him as a custodian, in relation to damages related to those aspects. For example, the user should not be a custodian in regard to software faults, when he has no access to the source code, and he is even forbidden to decode and modify it. To approach liability for failures concerning such aspects, the idea of custody is inadequate: either we extend it, so that it covers also the role of producers, designers and developers (this is a direction that has been taken by some legal systems, in particular the French one), or we may supplement it by appealing to different branches of the law (e.g. product liability, consumer protection, etc.).

A second issue, still common in regard to other technological entities, concerns the extent of the liability of the custodians. Two views are possible in regard to this form of liability. According to the first idea, a custodian is liable only when, and to the extent that, he has negligently omitted to control the thing. However, it may be very difficult to identify a lack of control in the user of a SA, since SAs have the capacity to act beyond the control of their users, and in ways that the latter could not foresee. If we follow this approach, then we have to conclude that in many (and maybe in most) cases, nobody would be liable for damages caused by SAs. Consequently

all Internet users would have to take the risk of supporting possible losses, as a consequence of the behaviour of SAs belonging to others. This may contribute to undermining trust in the net, and, considering the difficulty of proving lack of control, may provide little incentive for responsible use of agent-based technologies.

An alternative view would consist in assuming that the custodian of a SA is always liable for any damage caused by the SA, regardless of his violation of a duty of care, i.e., placing strict liability upon the custodian. This would allow economic losses caused by SAs to be transferred from the damaged persons to the custodian. This solution may seem harsh in regard the custodian (the user), who would face an unpredictable liability, even for events that are beyond his control. However, some assistance to the allocation of liability may be provided by the standard usually suggested by the law and economics school: put liability on the shoulders of the person who can more cheaply prevent damages (or insure against them). According to this criterion we would need to put liability, according to the type of problem that caused the damage, either on the developer, or on the owner, or the user of the SA. Moreover, we may also take into account contributory negligence on the part of the damaged person.

So far, we have remained within the boundaries of well-known legal problems, where the issues related to SAs are not so different from those one may address in relation to other technological objects.

The “new” issues that we need to solve are related to the feature of SAs we introduced above, the fact that as we said, they are cognitive tools. We need to consider whether the cognitive states of a SA are relevant to establishing and circumscribing the liabilities deriving from a damage that has been caused by that SA. Note that this does not amount to asking whether a SA is legally responsible, and even less to ask whether it is morally responsible. This is irrelevant to us, since by “liable” (legally, morally, or whatever) we just mean “obliged to pay compensation”, and the only liability we are considering is the user’s liability. We will argue that, even if only the user is liable (responsible) the fact that the user’s liability may depend upon the cognitive states of his SA, differentiates SAs from other things or tools, and justifies drawing analogies to vicarious liability for human actions.

This aspect comes to the fore when we have to decide what events have been properly “caused” by a SA. This is very important if we adopt a strict (no-fault) liability for the user, since following this approach, the fact that a SA caused damage would be both a necessary and sufficient ground for the

liability of the user.

Consider for example the case of a SA sending an innocent message to a computer system (“price offered Euro 75”), and assume that this message initiates a process leading the addressed system to crash, due to a fault of that system. Assume that this message was a necessary condition for the crash to happen (without the message the crash would not have occurred). To allocate liability we need to answer the following question. Did the message really “cause” the crash (so that the user of the SA, being its custodian, will have to pay damages), or was some defective procedure of the addressee system the real “cause”, and the message only provided the occasion for the internal fault to operate? One criterion to limit causality is the idea of “normal” or “adequate” causality (we cannot discuss here the issue of causality, which is indeed one of the most debated legal concepts): one event only “causes” those effects that normally follow from it: an exceptional effect, due to exceptional concurring factors (such as the system’s malfunctioning, in our example) does not really count as being caused by the event.

The limitation of causality to “normal” effects may lead to apparently absurd results in the case of damages which are intentionally (deliberately) produced, even when the intention at stake is the SA’s intention rather than the user’s.

Let us consider two different hypothetical cases. In the first, the SA sending the message knew of the existence of the faulty procedure, and sent the message exactly in order to produce the crash. In the second, the SA sent the message in good faith, in order to make a purchase offer. The two hypotheticals are identical in regard to the external behaviour of the involved SAs, which consists in performing a normally innocent action: sending the message “price offered Euro 75”. The only difference lies in the reasons that motivated the concerned SAs to send the message in the two hypotheticals.

Now, either the two hypotheticals have to be treated in the same way, or they have to be distinguished according to the different cognitive states that the two SAs had (assume, for the sake of the example, that a way is available of ascertaining these cognitive states). If the two cases have to be treated in the same way, there should be a verdict of non-liability in both cases. But this would provide an incentive for constructing SAs that tend to exploit defects of other systems, since the user of such SAs will never be liable for normally innocent (though intentionally malicious) behaviour of his SAs. Therefore, it seems that we need to conclude for the need to differentiate the legal discipline of the two hypotheticals: damage intentionally caused by

the SA should determine liability of its user, even when the damage was due to exceptional circumstances (especially when such circumstances were known to the SA), while damage unintentionally caused should not produce this result under the same circumstances. More generally, following the latter approach, a user would be liable for damages that have been “deliberately” produced by his SA (those damages that the SA intended to realise, or that it foresaw as being effects of its action), and for damages the SA produced as a consequence of violating duties of care concerning the activity it was performing (damages the SA should have foreseen and avoided).

Note that the idea that a SA should respect duties of care does not imply that the SA is responsible (in the sense of being liable to punishment) for the violation of those duties. It only implies that if the SA does not behave as those duties of care require (if it fails to anticipate the likely effects of its behaviour, or to act accordingly to such anticipation, or to use appropriate cautions), then the SA has been faulty (as a cognitive device), so that its owner should be liable, as any user (or owner) of a faulty machine. On the other hand, if the SA used all care objectively required by the activity being performed, then the user should not be liable for damage resulting from the activity of his SA, since in such a case the SA has been functioning perfectly well, so that liability cannot be placed upon its owner (unless this is a situation where strict liability would apply to actions by the user himself). Moreover, even the idea that damages should be put upon the person who could most cheaply avoid them is consistent with the idea intentions of the SA may condition the user’s liability. If an agent deliberately caused a damage, then this implies that this damage could be cheaply avoided: the user could easily have constrained the behaviour of his system in such a way that it would refrain from taking such malicious initiatives.

So, it seems that the guardian’s liability for the action of a SA cannot be grounded only upon the fact that a damage could be foreseen according to the “normal” laws of nature (or of technology). We need rather to consider whether the SA intentionally or negligently produced the damage. If we have indeed to draw this conclusion, then the liability of the user of a SA would be similar, rather than to liability of a custodian of a thing, to vicarious liability (the liability of the employer for the employee). This form of liability is not based upon the fact that the employer could foresee the behaviour of the employee, but rather on the fact that the employee accomplished a tort, when acting in the course of the employment.

Note that the relevance of the SA’s cognitive states is the only reason why

one may consider assimilating the relation of a SA to its user to the relation of an employee to his employer. This has nothing to do with labelling a SA as a person, or as a legal or moral subject.

## 9 SAs and contracts

On line contracting (and bargaining) is already a very important application area for agent-based technologies (on the legal aspects of contracts made by SAs, cf. Lerouge [Ler00], Weitzenboeck [Wei01a], DeMiglio et al [DMORS02]). This is confirmed by the fact that some legislatures have already shown some interest for this domain. In particular, the US Uniform Computer Information Transactions Act (UCITA) establishes some rules that specifically concern electronic agents. UCITA defines an agent as “a computer program or electronic or other automated means, used independently to initiate an action, or to respond to electronic messages or performances, on the person’s behalf without review of action by an individual at the time of the action or response to the message or performance”, and (in section 107 (d)) affirms that “a person that uses an electronic agent that it has selected for making an authentication, performance or agreement, including manifestation of assent, is bound by the operations of the electronic agent, even if no individual was aware or reviewed the agent’s operations or the results of the operations.” Similar indications are contained in the above mentioned Unicitral document, which suggests that a new convention on electronic contracting should make explicit allowance for the validity of contracts made through automatic agents, even with no human control, and affirms that the user is responsible for the agent’s declarations, even when he has not wanted those declarations.

Also in regard to contracts made by SAs, we will concern ourselves with the most mundane applications. We will not even raise the issue whether self-interested SAs populating the Internet may be viewed as autonomous subjects, endowed with legal and moral subjectivity, who can enter into legally binding agreements in their own name. We will confine ourselves to SAs charged with negotiating and concluding contracts in the name and in the interest of their users. These are the hypotheses when the SA is buying or selling things in the name or on the account of its user, and more generally when it is performing activities pertaining to the formation of contractual offers or to their acceptance in the name of the user. We will argue that to

approach such contexts coherently we need to take seriously the idea of a SA having cognitive states, which may be relevant to the law.

First of all, we need to reject the view that SAs only transmit contracts prepared by their user (or programmer). This view is incompatible with the fact that neither the user nor the programmer are in such a condition to fully anticipate the contractual behaviour of the SA in all possible circumstances, and therefore to “want” the contracts which the SA will conclude. Even when the user is in the condition of making such a forecast, he cannot be required to do it, since, as we observed above, this would contradict the very reason for using an SA: delegating cognitive tasks, as the acquisition of knowledge and its use in deliberation. Therefore, the fact that the effects of a contract made by a SA fall upon the user is not explained by the fact that the user foresaw the behaviour of its SA, or could foresee it, or even ought to have foreseen it (as affirmed, for example, by Finocchiaro [Fin02]). It is true that the intention of the user (as recognised by the counterparty) provides the ultimate justification for the effectiveness of the contracts made by his SA, but this is his intention to entrust the SA with the task of entering into certain transactions in his name, performing the cognitive processes that are required for making those transactions.

The admission that the user does not have (and cannot be required to have) any cognitive state directly concerning the individual contracts made by his SA (no intention or wish that those contracts be made, nor any knowledge of their terms and preconditions) leads us into a difficult dilemma with respect to contracts made by SA (and in general by automatic systems, on which cf. for all Allen and Widdison [AW96]). We need to decide which of the following views to adopt:

1. Those contracts are not accompanied by any relevant cognitive states (they are exchanges without agreement, as an Italian jurist recently said (Irti [Irt98]), to be considered from a purely behaviouristic perspective. Therefore having adopted the decision of making a contract, possessing information that appropriate circumstances obtain, or believing that the counterparty has certain cognitive states, should be irrelevant to the effects of those contracts.
2. Those contracts are characterised by the cognitive states that are possessed (or may be attributed to) the SA making them. Therefore, the fact that a SA has formed a certain intention, or had certain beliefs,

at least when this was known to the counterparty, may impinge on the effects of the contract.

Consider for example the following hypothetical. Assume that a SA has been charged with the task of selling on line certain pieces of old jewellery according to their weight, age, and constitutional material they are built. Assume that the SA uses a database (prepared by an expert) where it can find a description of all items to be sold. Assume that item number 25 has been mistakenly classified as being a silver ring with a gold coating when it is gold ring. Assume that the SA offers to sell item 25 for the price of 20 euros, considering that this is a price appropriate for a silver ring, and that the counterparty accepts. Assume also that in the photograph of the ring available on line one could easily see the words “gold 18 K”. Will the contract be voidable since the decision to conclude it was based upon a mistake (the false belief that the ring was made of silver), and this was known to the counterparty, or will the contract be valid, since a SA cannot have any cognitive states, and therefore cannot make any mistake? In general, legal systems allow such contracts to be voided when the mistake was an important one and was recognisable to the counterparty. What will happen when, as in the case at hand, the mistake was committed by a SA? And what if the counterparty knew that the SA was making a mistake? And what if the SA’s mistake was induced by the counterparty, which, for example, provided a wrong input to the SA’s database?

One way to evaluate this situation is the behaviouristic approach, which requires refraining from any use of cognitive notions when dealing with SAs: any legal effect is directly linked to a specific observable behaviour, not to the cognitive states which may be inferred from observable behaviour. What matters is only the fact that certain data messages were sent, having a certain conventional meaning. In the example we considered above, since appropriate offer and acceptance messages were sent, the conclusion will be for the validity of the contract. As this example shows, the behaviouristic approach, though being sensible to a certain extents (as when the parties may have agreed to give a certain pre-established effects to certain actions of their systems), may lead to absurd results, and cannot provide the flexibility of an approach based upon intentional notions. The problem with a behaviouristic approach is that it is impossible to specify in advance what observable behaviour will correspond to a certain cognitive state (e.g. to the belief that something is the case, or to the intention of producing a certain result). If

we directly and indefeasibly link to specific observable behaviour the legal effects that should follow from having certain cognitive states, then those effects will sometimes follow even when those cognitive states are absent, and they may not follow when they are present. Such an approach lends itself to being opportunistically exploited, as when one SA is tricked into sending a certain message, though not having any “intention” of performing the corresponding communicative action. Moreover, an unpleasant consequence of this approach would be the need to duplicate any legal notion, for the purpose of applying it to computer systems, in order to provide behavioural equivalents of the intentional notions used for governing human interaction (in Italian law, the idea of computer contracts as a new way of creating legal relations, different from human contracts, has been advanced by Giannantonio [Gia97]).

The alternative possibility consists in taking seriously the intentional stance, in attributing cognitive states to SA’s (as to other computer systems), and considering whether having those states (and being attributed them) may make a difference in the legal effects of the behaviour of such systems. This would amount to assuming that the intentional stance, being the only viable way of approaching a (complex and autonomous) system, can also be given legal recognition.

According to this approach the counterparty of a SA, when the behaviour of the SA provides adequate clues, can interpret its contractual declaration by attributing to the SA the corresponding intentions (e.g., the intention to sell the ring), and the epistemic states that are presupposed by such intentions (e.g., the belief that the ring is made of silver). Usually, the fact that a SA makes a certain declaration, in appropriate circumstances, would be a sufficient clue to the SA’s intentional states. Moreover, whenever the counterparty reasonably believed that the SA had such intention, on the basis of the SA’s behaviour, the fact that such intention did not really exist would usually be irrelevant (according to the principle of the protection of justified reliance). Therefore, in the vast majority of cases a behavioural approach and an intentional approach would lead to the same practical results.

However, when the SA (though sending a certain data message) has no intention of making a contract (for example, the SA produces the message to comply to somebody’s request of forwarding a sequence of words), and the counterparty is aware of that (or should be aware, given the circumstances of the case), no contract will be concluded. It also implies that defects in the cognitive processes of a SA, as impairment to the formation of the

contractual volition, should have legal implications which are similar to the so called defects of will (mistake, deceit, duress).

In general, the fact that the content of the contract is determined (also) by the SA, does not exclude that the rights and the duties created through the contract should fall upon the user. This is exactly what the user wants, when he delegates the formation of the contract to his SA. So, the intention of the user (as recognisable by the counterparty) to delegate the formation of the contract to the SA's cognition is the ground on which the contract is non-repudiable by the user, though the user has not wanted the specific content of the contract concluded by the SA. The rights and obligations issuing from the contract will fall upon the user, not because he wanted those contents, but because he has chosen to delegate to his SA the formation of contracts in his name. Cognition by the SA complements cognition by the user, according to the intention of the user, and should be treated, in principle, in the same way. This leads us to assimilate the situation of the user of a SA to the situation of a person handing over the conclusion of a contract to a human agent (in Italian law, this idea has been advanced by Borruso [Bor88] in regard to computer-made contracts in general). What the two situations have in common, which distinguishes them from the situation where one uses a (mechanical or human) means of transmission, is cognitive delegation, i.e. the decision to entrust the formation of the content of a contract and the decision whether to conclude it or not (though within pre-established objectives and constraints), to someone (or something) else's cognition.

This perspective excludes that each determination of the SA necessarily is (or should be) a determination of his user: when it is necessary to establish who wanted what, we need to examine which contents of the contract were pre-established by the user, and which ones were determined by the SA. Consequently, when one has to establish whether the conclusion of the contract was due to deceit (so that the contract can be voided), in regard to the elements which were determined by the user, one has to look whether the user was cheated, but in regard to the elements which were determined by the SA, one must look whether the SA was induced into error. As to the effects of a mistake, one has to consider that a mistake will in general impact upon the validity of a contract, only if it is recognisable to the counterparty. This circumscribes the effect of mistakes (false beliefs, or false epistemic states), both when they are made by the user and when they are made by a SA. If a SA's mistake is not recognisable to the counterparty, the contractual declaration made by the SA (within the domain where the SA reasonably appears

to be acting within the delegation of the user) will bind its user.

It seems that in this regard the model of commercial agency (and representation) may give appropriate clues. In fact, according to this model, malfunctioning in the deliberation process of the SA (and consequently its mental states) impacts on the validity of the contract. This view seems to be compatible with the following statement, included in the above mentioned Unicitral document:

the Working Group was of the view that, while the expression “electronic agent” had been used for purposes of convenience, the analogy between an automated system and a sales agent was not appropriate. Thus, general principles of agency law (for example, principles involving limitation of liability as a result of the faulty behaviour of the agent) could not be used in connection with the operation of such systems. The Working Group reiterated its earlier understanding that, as a general principle, the person (whether a natural person or a legal entity) on whose behalf a computer was programmed should ultimately be responsible for any message generated by the machine (A/CN.9/484, para. 107). As a general rule, the employer of a tool is responsible for the results obtained by the use of that tool since the tool has no independent volition of its own. However, an “electronic agent”, by definition, is capable, within the parameters of its programming, of initiating, responding or interacting with other parties or their electronic agents once it has been activated by a party, without further attention of that party.

More exactly, we must agree on the proposition that it would be inappropriate to automatically transfer to users of SAs any rule applicable to principals in regard to their sales agents (since in particular, sales agents may be liable on their own). However, it seems wrong to affirm that a necessary connection exists between responsibility (liability to pay damages) of the user, and the lack of volition in its tools: when the user is using a cognitive tool, as a SA, he may well be liable also for actions that his tool has intentionally accomplished (a doubt in this regard seems to emerge in the last phrase in the citation above).

Therefore, the fact that in regard to both commercial agents and to SAs cognitive delegation is at issue, may justify drawing useful analogies, since the same rationale may apply to both situations. Consider for example the issue

of the time of the conclusion of a contract, in regard to mobile SAs. A mobile SA may operate remotely without interacting with its user and with the computer system of the latter (consider for example, a SA inhabiting a mobile device). If the SA were only a means for the user to communicate with other parties, contracts concluded through the SA would be finalised only when the acceptance of the other party reaches the user (or at least the computer system of the latter), since (at least according to Italian law) a contract is concluded when acceptance reaches the offeror. This may cause difficulties in some applications of SAs. Consider for example a mobile SA, which moves into a financial marketplace, and then proceeds to buy and sell stock. Is it reasonable to assume that the contracts concluded through the SA are finalised only when they reach the computer of the user? This would preclude mobile SAs from selling what they have just bought, before communicating the purchase to their user (since communication to the user is necessary to perfect the previous exchange), and therefore from engaging in effective on-line trading. Again, the model of representation (sales agent) may provide the right clue to approach agent-based negotiation: the contract concluded through a representative is usually finalised when acceptance reaches the representative.

## 10 Further issues

Let us now shortly consider some further issues in the law of SAs (for an introduction to those issues, cf. Weitzenboeck [Wei01b]).

- SAs and consumer protection
- SAs and legal personality
- SAs and intellectual property
- SAs and privacy protection
- SAs and right to information

### 10.1 SAs and consumer protection

In discussing SAs and consumer protection, I will refer in this regard to Italian law, but this law is similar to the law of many other countries (see,

for example vanHaentjens [vH02], Rossato [Ros02]). Consumer protection is to a large extent achieved by making certain contractual terms (which would impair the position of the consumer) ineffective. In particular, the Italian civil code has a special rule (art. 1341) concerning contracts made through standardised forms: some terms, which are likely to impair the position of one party, are not effective unless they are singularly approved in writing by that party, which means separately signed (this concerns for example, arbitration clauses, or clauses excluding liability of the counterparty). According to art. 1469 bis of the Italian civil code (which implements European legislation, and therefore establishes a discipline which is largely common to other EU countries) a larger list of contractual terms, which are likely to impair the position of the customer, are ineffective in contracts between a professional operator and a consumer. Such terms will only be effective when they have been the subject of a specific negotiation. Now, if both the professional operator and the customer were acting through SAs (possessing an electronic signature), then the SA's customer could singularly sign each clause which needs to be signed, and both SAs could singularly negotiate each clause that needs to be negotiated. Consider for example, how a customer (or its SA) could negotiate with the seller's SA, and accept a change in the determination of the competent judge, or in the applicable law, or a limitation of the seller's liability, in exchange for a reduction of the price.

More generally, the use of SAs, by eliminating transaction costs (negotiation through SAs would be practically costless) would make irrelevant every law establishing contractual terms which can be derogated by the parties, since those terms could always be substituted by the result of a negotiation. This implies that SAs may make irrelevant, as far as the substance of economic relations is concerned, any rule attributing renounceable rights, according to the famous theory of Coase [Coa60], who argued that with no transaction costs, economic efficiency alone decides the allocation and the use of resources .

## 10.2 SAs and legal personality

Up to this point, we have considered how users may acquire rights and duties through the activity of their SAs. This result only requires the assumption that SAs have a power to produce rights and duties for others. Now we are going to investigate whether SAs may also have their own rights and duties, that is whether they may be legal persons in the sense of non human bearers

of rights and duties.

If a SA were considered a legal person, then it would be able to enter into contracts in its own name, and to acquire rights and duties. At a later time, the SA may transfer these rights to its user, or even transfer them to a third party (consider for example a SA who acts as an on-line trader, buying certain commodities and reselling them at a higher price, if it can find a buyer). These rights would be included in the patrimony of the SA, until the subsequent transfer takes place. We may imagine that this patrimony would be started by the user, who transfers to the SA an amount of money (obviously, electronic money), to be used in on-line transaction. This fund would represent a warranty for the counterparties, who would need to know its amount before finalising a contract with the SA.

What distinguishes transactions where the SA acts on its own, from the ones where it represents its user, is that in the first type of transactions, the counterparties would not know on whose behalf the SA is acting. If the SA does not fulfill its obligations, we may then imagine that the creditors of the SA would first “sue the agent”, that is try to be compensated with the money in the SA’s fund. Only if the patrimony of the SA were insufficient, would they try to discover who is the user-owner of the SA, and try to get compensation from him.

Having the capacity of bearing rights and duties does not yet provide full legal personality. This would require, as for legal persons, a complete separation of the patrimony of the SA from the patrimony of its users. If a SA had full legal personality, then the creditors of the SA could only sue the SA, in order to be satisfied with what is included in the SA’s patrimony. The user-owner of the SA would have no liability (when the SA made a contract in its own name), beyond the amount he has transferred to the SA’s fund.

The personification of SAs would reassure their owners-users, since they would know that they would not suffer any loss beyond the amount of money they have transferred to the SA’s patrimony. However, conferring legal personality on SAs might create various difficult legal problems. First of all, SAs do not have an established physical location. At what residence or domicile would the unsatisfied creditor sue a SA? Secondly, SAs may disappear, definitely (being cancelled) or temporarily (being registered on an inaccessible storage device), they can divide themselves into the modules that they include, and they can multiply themselves into indistinguishable copies. How is it possible to identify precisely the entity that holds the obligations and rights of the SA? Thirdly, behind the screen of a personified SA, various

abusive practices may take place (e.g., the user may take away the money in the SA's fund, for example by simulating a sale to the SA, and then let it go bankrupt, and default in its obligations)

Some of these problems can possibly be solved through some legal artifices: for example the residence or the domicile of a SA may be the address of the bank where the SA's fund is deposited, the acts attributed to the SA would be those which are signed with the SA's digital signature, special controls on personified SAs could be devised, etc.

However, giving legal personality to SAs does not seem necessary or even opportune. An easier and less risky way for the SA to make contracts without revealing the name of its user, and to limit the liability of the user (at least to some extent) is available. This consists in creating companies for on line trading, which would use SAs in doing their business. Such SAs would act in the name of a company, their will would count as the will of the company, their legally relevant location would be the company's domicile, and creditors could sue the company for obligations contracted by those SAs. The counterparties of a SA could then be warranted by the capital of the company and by the legal remedies available against defaulting commercial companies.

Finally, let us observe that there are no obstacles to creating special normative systems - for example, the regulation of an on-line marketplace - that directly govern the activities of SAs. Within such normative systems, SAs may hold normative positions (rights and duties) and have a full subjectivity. Those positions would not be recognised directly in the legal system (SAs will still have no legal rights and duties), but nevertheless, they could have some legal consequences. For example, the owner or the user of the SA may be legally obliged to pay a penalty if the SA violates the rules of the marketplace. So, though we should not see SAs as being addressees of legal norms (see, however, Taddei Elmi[TE90] and Karnow [Kar94]), it makes sense to speak of normative SAs, and to design SAs having the cognitive competence for adopting and complying with norms (cf. Castelfranchi et al [CDCT99], Artosi [Art02], Boella and Damiano [BD02], Brazier et al [BKOW02], Gelati et al [GRS02]).

### **10.3 SAs and intellectual property**

Here we will not even mention the many important intellectual property issues which are related to the use of SAs (for a discussion of this topic,

cf. Bing and Sartor [BS02], and Bing [Bin02]). Let us just observe that in regard to the issue of SAs and intellectual property, one may distinguish three aspects: (a) the protection of the SA itself; (b) the protection of the results of the activity of the SA (when engaged in searching and processing information); (c) the protection of the information sources accessed by the SA. Concerning the first issue, we must consider that SAs include innovative technologies. Those technologies may possibly be patented (according to the approach already adopted in the USA and currently debated in Europe). In particular, we need to consider that SAs and multi-agent systems can implement a vast number of business methods. If those methods can be patented, when implemented in a computer system (as it is now the case in the USA), then there is a prospect for patenting the structures of agent-based societies and the patterns of agent interactions. This raises very important issues for the evolution and the commercial exploitation of agent technologies: patents provide a powerful incentive, but may also unduly constrain research and applications.

Concerning the second issue (information collected and processed by SAs) the basic reference, at least in Europe, is the protection of databases (according to directive 96/9/CE, and national legislations implementing it). In this regard, we need to establish when data collected and organised by a SA can be qualified as a database. An issue for the future is whether works realised by electronic artists may have the level of creativity required for copyright protection. In regard to the creation of an SA-author, should we apply the same criteria that we apply for human-authors? What rights belong to the SA's creator, what to the user or the owner? And what about the moral rights of the author?

## 10.4 SAs and privacy

In the privacy domain (cf. Borking et al [BVES99], Bygrave [Byg01], Villecco [Vil02]), there are two main issues to be addressed. On the one hand, SAs may violate people's privacy, by collecting data concerning individuals, and processing it contrary to the standards of data protection. An interesting issue concerns how the processing of such data can be limited to legitimate purposes, previously communicated to the concerned individuals, as required by European legislation. This constraint seems hard to implement in regard to SAs, given their autonomy. On the other hand, SAs can be the victims of privacy violations. In particular, a SA may contain a profile of its user, in

order to be able to act on the user's interest (the credit card number of the user, his electronic signature, a description of his needs and tastes, a record of his previous purchases, etc.). Third parties accessing this data would violate the privacy of the user. One may wonder whether there is a sense in which also the privacy of the SA can be protected. This concerns, for example, the privacy of the cognitive states of the SA (which do not need to be cognitive states of its user, as we have observed above), the knowledge of which may provide an unfair advantage to the counterparty, even when those cognitive state do not correspond to a cognitive state of the user. For example, access to negotiation strategies recorded in the SA's memory may give a decisive advantage in negotiations with the SA (assume for example, that the seller comes to know the maximum price that the buyer-SA is able to pay).

## 10.5 Right to information

A further issue concerns the use of information agents in accessing data pertaining to social and political issues. It has been argued that SAs could filter the available data, and exclude some information from being accessed by the public. This would prevent the formation of an informed public opinion, and so create an impediment to democratic debate. This concerns, in particular, the fact that an important input to the formation of one's political opinion consists in the unrestricted exposure to information relevant to political issues (e.g. information about poverty, deprivation, etc.), even to information that one may prefer not to see, for the sake of one's peace of mind (for a discussion of this issue, cf. Lessig [Les99], 164 ff, and Sunstein [Sun01]). On the other hand, however, one may argue that information agents may be an important instrument of deliberative democracy, by allowing individuals to access information concerning political issues they are interested in, information that would be irretrievable without adequate search tools. Forbidding the use of information agents would also be an inadmissible limitation of the freedom of information. A legitimate use of information agents seems to require that the criteria they use in selecting information are made explicit to their users (in an understandable form) and that there is a decentralised and uncontrolled provision of information agents. However, it remains true that SAs may allow their users to effectively shield themselves against unwanted information, in a way that may have a negative impact on democracy.

## 10.6 Future scenarios

There is no possibility of considering here many further legal issues which are related to SAs, such as their use in virtual enterprises (cf. Cevenini [Cev02]) or in on-line dispute resolution (cf. Chiti and Peruginelli [CP02], Gouimenou [Gou02]). We will not even try to consider here the legal issues that may arise in a distant future as described by Kurzweil [Kur99], where the distinction between humans and SAs becomes uncertain, as a consequence on the one hand of installing hardware and software prostheses in human beings and on the other hand of creating more and more complex virtual agents. This author imagines a progressive intertwining of reality and cyberspace, that would lead (in less than a century), to the possibility of passing from one dimension to the other: human individuals could have an electronic existence (so obtaining, among other things, immortality) and SAs could be embodied in physical and even biological structures. What will happen, for example, to inheritance law, when individuals would be souls, moving from one substrate to the other? What will be the legal relationship between the various embodiments of one individual?

Rather than considering such improbable issues, we need to mention the most radical critiques on the use of SAs. It has been affirmed that using SAs would imply that we renounce our human and social competence. Once we admit SAs in typically human relations, we would need to adapt to the logic of impoverished interactions, in which it would hard for us to compete with our digital assistants. Moreover, using SAs as intermediaries for accessing goods and experiences would contribute to compromising the chance of establishing authentic relationships with other persons.

The latter argument needs to be taken seriously. However, we need to consider that the agents' model does not identify a set of specific software products (we may even argue that there are yet very few real SAs are commercially available). Rather it is a comprehensive paradigm for computing, a paradigm which may lead to very different applications. We need to ensure that the use of this paradigm will provide us with trusted electronic intermediaries without forcing us to renounce our capacity for decision and interaction, and that it will increase rather than diminish security and trust. The realisation of these objectives depends on various factors (political, economic, and technological ones). However an essential precondition is also the definition of an appropriate legal framework.

## 11 Acknowledgements

This work has been supported by the EU IST FET UIE project ALFEBIITE (IST-1999-10298), and this support is gratefully acknowledged. I would particularly like to thank the partners in this project for providing the context for the current work. However, the author himself is solely responsible for any opinions or mistakes contained in this article.

## References

- [Art02] Alberto Artosi. On the notion of an empowered agent. In *Proceedings of LEA 2002: Workshop on the Law of Electronic Agents*, pages 00–00. CIRSFID, Bologna, 2002.
- [AW96] T. Allen and R. Widdison. Can computers make contracts. *Harvard Journal of Law and Technology*, 9:25–52, 1996.
- [BD02] Guido Boella and Rossana Damiano. A game-theoretic model of third-party agents for enforcing obligations in transactions. In *Proceedings of LEA 2002: Workshop on the Law of Electronic Agents*, pages 111–121. CIRSFID, Bologna, 2002.
- [Bin02] Jon Bing. Electronic agents and intellectual property law. In *Proceedings of LEA 2002: Workshop on the Law of Electronic Agents*, pages 00–00. CIRSFID, Bologna, 2002.
- [BKOW02] Frances Brazier, Onno Kubbe, Anja Oskamp, and Nick Wijngaards. Are law abiding agents realistic? In *Proceedings of LEA 2002: Workshop on the Law of Electronic Agents*, pages 151–157. CIRSFID, Bologna, 2002.
- [Bor88] Renato Borruso. *Computer e diritto: Problemi giuridici dell'informatica, Vol. 2*. Giuffrè, Milano, 1988.
- [BS02] Jon Bing and Giovanni Sartor. Lovely rita: A scenario. Deliverable, ALFEBIITE (IST-1999-10298), 2002.
- [BVES99] J. J. Borking, B. M. A. Van Eck, and P. Siepel. *Intelligent software agents and privacy*. Registratiekamer, The Hague, 1999.

- [Byg01] Lee A. Bygrave. Electronic agents and privacy: A cyberspace odyssey 2001. *International Journal of Law and Information Technology*, 9:275–294, 2001.
- [CDC00] Robert Cummins and Denise Dellarosa-Cummins. *Minds, Brains, and Computers: The Foundations of Cognitive Science*. Blackwell, London, 2000.
- [CDCT99] Cristiano Castelfranchi, F. Dignum, M. J. Catholijn, and J. Treur. Deliberative normative agents: Principles and architecture. In *Proceedings of ATAL 1999*, pages 364–378, 1999.
- [Cev02] Claudia Cevenini. Agents in the virtual enterprise: Some legal notes. In *Proceedings of LEA 2002: Workshop on the Law of Electronic Agents*, pages 59–64. CIRSFID, Bologna, 2002.
- [CF03] Cristiano Castelfranchi and Rino Falcone. Socio-cognitive theory of trust. In Jeremy Pitt, editor, *The Open Agent Society*. Wiley, London, 2003. (Forthcoming).
- [CGV02] Claudia Cevenini, Jonathan Gelati, and Alessandra Villecco. *Proceedings of LEA 2002: Workshop on the Law of Electronic Agents*. CIRSFID, Bologna, 2002.
- [Chu95] P. M. Churchland. *The Engine of Reason, the Seat of the Soul: A philosophical Journey into the Brain*. MIT, Cambridge, Mass., 1995.
- [Chu00] P. M. Churchland. Eliminative materialism and the propositional attitudes. In R. Cummins and D. Dellarosa-Cummins, editors, *Minds, Brains and Computers: The Foundations of Cognitive Science*, pages 500–512. Blackwell, London, 2000. (First published 1981.).
- [Cic01] Antonio Cicu. Gli automi nel diritto privato. *Il Filangeri*, 8:1–30, 1901.
- [Coa60] R. Coase. The problem of social cost. *The Journal of Law and Economics*, 3:1–44, 1960.

- [CP02] G. Chiti and G. Peruginelli. Artificial intelligence in alternative dispute resolution. In *Proceedings of LEA 2002: Workshop on the Law of Electronic Agents*, pages 97–104. CIRSIFID, Bologna, 2002.
- [Dav98a] Martin Davies. The philosophy of mind. In A. C. Graylin, editor, *Philosophy 1: A Guide Through the Subject*, pages 250–335. Oxford University Press, Oxford, 1998.
- [Dav98b] J. R. Davis. On self-enforcing contracts, the right to hack, and wilfully ignorant agents. *Berkley Technology Law Journal*, page 1148, 1998.
- [Daw89] Richard Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, 1989.
- [Den89] Daniel C. Dennett. *The Intentional Stance*. MIT, Cambridge, Mass., 1989.
- [Den91] Daniel C. Dennett. *Consciousness Explained*. Little Brown, Boston, Mass., 1991.
- [Den96] Daniel C. Dennett. *Darwin’s Dangerous Idea: Evolution and the Meanings of Life*. Penguin, London, 1996. (First published 1995.).
- [Den97] Daniel C. Dennett. *Kinds of Minds: Towards an Understanding of Consciousness*. Basic Books, New York, N. Y., 1997.
- [DH87] Daniel C. Dennett and John C. Haugeland. Intentionality. In R. L. Gregory, editor, *The Oxford Companion to the Mind*, pages 383–386. Oxford University Press, Oxford, 1987.
- [DMORS02] F. De Miglio, T. Onida, F. Romano, and S. Santoro. Electronic agents and the law of agency. In *Proceedings of LEA 2002: Workshop on the Law of Electronic Agents*, pages 23–32. CIRSIFID, Bologna, 2002.
- [Dre86] Fred Dretske. Misrepresentation. In R. J. Bogdan, editor, *Belief: Form, Content and Function*, pages 17–36. Oxford University Press, Oxford, 1986.

- [ET00] Gerald M. Edelman and Giulio Tononi. *A Universe of Consciousness*. Basic Books, New York, N. Y., 2000.
- [Fin02] Giusella Finocchiaro. The conclusion of the electronic contract through “software agents”: A false legal problem? brief considerations. In *Proceedings of LEA 2002: Workshop on the Law of Electronic Agents*, pages 75–80. CIRSFID, Bologna, 2002.
- [Fod83] J. Fodor. *The Modularity of Mind*. MIT, Cambridge, Mass., 1983.
- [Gia97] Ettore Giannantonio. *Diritto dell’informatica*. Giuffrè, Milano, seconda edition, 1997.
- [Gou02] J. Gouimenou. E-arbitration-t ©: An alternative dispute resolution for SMEs. In *Proceedings of LEA 2002: Workshop on the Law of Electronic Agents*, pages 105–110. CIRSFID, Bologna, 2002.
- [GRS02] Jonathan Gelati, Nino Rotolo, and Giovanni Sartor. Normative autonomy and normative co-ordination: Declarative power, representation, and mandate. In *Proceedings of LEA 2002: Workshop on the Law of Electronic Agents*, pages 133–150. CIRSFID, Bologna, 2002.
- [Hau00] John Haugeland. Semantic engines: An introduction to mind design. In R. Cummins and D. Dellarosa-Cummins, editors, *Minds, Brains, and Computers. The Foundations of Cognitive Science*, pages 34–50. Blackwell, London, 2000. (First published 1981.).
- [Irt98] Natalino Irti. Scambi senza accordo. *Rivista trimestrale di diritto e procedura civile*, pages 347–364, 1998.
- [Jon00] Andrew J. Jones. On the concept of trust. 2000.
- [Kar94] C.E.A. Karnow. The encrypted self: Fleshing out the rights of electronic personalities. *The John Marshall Journal of Computer and Information Law*, 13:1–16, 1994.

- [Kur99] Ray Kurzweil. *The Age of Spiritual Machines*. Orion, London, 1999.
- [Ler00] J. F. Lerouge. The use of electronic agents questioned under contractual law: Suggested solutions on a european and american level. *The John Marshall Journal of Computer and Information Law*, 18 (2):430–00, 2000.
- [Les99] Lawrence Lessig. *Code and Other Laws of Cyberspace*. Basic Books, New York, N. Y., 1999.
- [Mil01] Ruth G. Millikan. Biofunctions: Two paradigms. In R. Cummins, A. Ariew, and M. Perlman, editors, *Functions in Philosophy of Biology and Philosophy of Psychology*. Oxford University Press, Oxford, 2001.
- [Nag74] Thomas Nagel. What is like to be a bat. *Philosophical review*, 83:435–450, 1974.
- [Noz93] Robert Nozick. *The Nature of Rationality*. Princeton University Press, Princeton, N. J., 1993.
- [PE77] Karl R. Popper and John Eccles. *The Self and Its Brain*. Routledge, London, 1977.
- [PMC01] Jeremy Pitt, Abe Mamdani, and P. Charlton. The open agent society and its enemies: A position statement and a programme of research. *Telematics and Informatics*, 18 (1):67–87, 2001.
- [Ros02] A. Rossato. “Stop the bot!”: Trespass to chattels in cyberspace. In *Proceedings of LEA 2002. Workshop on the Law of Electronic Agents*, pages 159–172. CIRSIFID, Bologna, 2002.
- [Ryl49] G. Ryle. *The Concept of Mind*. Hutchinson, London, 1949.
- [Sar02] Giovanni Sartor. Intentional concepts and the legal discipline of software agents. In Jeremy Pitt, editor, *Open Agent Societies: Normative Specifications in Multi-Agent Systems*. Wiley, London, 2002. (Forthcoming).
- [Sea89] John R. Searle. Consciousness, unconsciousness, intentionality. *Philosophical topics*, 17:193–209, 1989.

- [Sea90] John R. Searle. Consciousness, explanatory inversion and cognitive science. *Behavioural and Brain Sciences*, 13:585–596, 1990.
- [Sea95] John R. Searle. *The Construction of Social Reality*. The Free Press, New York, N. Y., 1995.
- [Sun01] Cass R. Sunstein. *Republic.com*. Princeton University Press, Princeton (N.J.), 2001.
- [TE90] Giancarlo C. Taddei Elmi. I diritti dell’intelligenza artificiale tra soggettività e valore: fantadiritto o jus condendum. In L. Lombardi Vallauri, editor, *Il meritevole di tutela*, pages 685–711. Giuffrè, Milano, 1990.
- [vH02] O. van Haentjens. Shopping agents and their legal implications regarding austrian law. In *Proceedings of LEA 2002. Workshop on the Law of Electronic Agents*, pages 81–96. CIRSIFID, Bologna, 2002.
- [Vil02] Alessandra Villecco. Agent technology and on-line data protection. In *Proceedings of LEA 2002. Workshop on the Law of Electronic Agents*, pages 53–58. CIRSIFID, Bologna, 2002.
- [Wei01a] Emily Weitzenboeck. Electronic agents and the formation of contracts. *International Journal of Law and Information Technology*, 9(3):204 – 234, 2001.
- [Wei01b] Emily Weitzenboeck. Electronics agents: Some other legal issues. ECLIP 2nd Summer School, Palme De Mallorca. Available at: <http://www.eclip.org/summerschool/2nd/presentations>, 13 October 2001 2001.
- [YC02] Alexander Yip and Jim Cunningham. Some issues on agent ownership. In *Proceedings of LEA 2002. Workshop on the Law of Electronic Agents*, pages 13–22. CIRSIFID, Bologna, 2002.