Artif Intell Law (2009) 17:253–290
DOI 10.1007/s10506-009-9081-0

Cognitive automata and the law: electronic contracting and the intentionality of software agents

Giovanni Sartor

Published online: 9 October 2009

Abstract I shall argue that software agents can be attributed cognitive states, since their behaviour can be best understood by adopting the intentional stance. These cognitive states are legally relevant when agents are delegated by their users to engage, without users' review, in choices based on their the agents' own knowledge. Consequently, both with regard to torts and to contracts, legal rules designed for humans can also be applied to software agents, even though the latter do not have rights and duties of their own. The implications of this approach in different areas of the law are then discussed, in particular with regard to contracts, torts, and personality.

Keywords Digital agents · Representation · Delegation · Responsibility

1 Cognitive states of artificial agents

The 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 may happen in some cases (which are likely to become more frequent in the future) through software agents (SAs), i.e., digital entities capable of executing autonomously the mandates assigned to them.[1] We naturally tend to apply

[1] When speaking on general terms of an "agent", I shall refer to any entity capable of autonomous action in general, following the AI terminology. In the law, on the contrary, the term "agent" usually denotes someone who acts on behalf of another. Both meanings, however, are relevant, since I shall consider autonomously electronic entities acting on behalf of their users. For useful links to research projects and

G. Sartor (✉)
European University Institute, Florence and Cirsfid, University of Bologna, Bologna, Italy
e-mail: giovanni.sartor@eui.eu

also to artificial entities, and especially to SAs, the interpretative models we apply to humans. In particular, we tend to explain the behaviour of such entities by attributing them cognitive states (beliefs, desires, intentions etc.). Consequently, we tend to qualify their actions through legal notions presupposing the attribution of cognitive states. Consider, for example, the possibility that a computer system enters into a contract (intends to execute the contract and thereby realise the legal results stated in the contract), is the object of a fraud (is cheated), makes a mistake (has a false belief), harms somebody with malice (intentionally), etc.

Our tendency to attribute cognitive states to artificial systems, and to apply the consequent legal qualifications conflicts with the assumption that cognitive (mentalistic) concepts only apply to humans. This assumption would imply the necessity to undertake an extensive review of the existing legal notions in order to apply them to artificial systems. For this purpose we would need to eliminate from legal notions any connection with mental or spiritual attitudes: will (intention) needs to be removed from the notion of a contract, having false beliefs from the notion of a mistake, causing false beliefs from the idea of misrepresentation, malice from the preconditions of criminal liability ... However, it is dubious that this strategy may lead us to results appropriate to our needs. In fact, it would force the legislator and the jurist to make the following choice: either to eliminate any cognitive (mental) notion from the law, or to duplicate the characterisation of legally operative 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. Eliminating cognitive notions 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 cognitive states determining and accompanying an action are often decisive elements for its evaluation). Duplicating operative facts adds unnecessary complexity to the legal system and induces incoherence, by providing for different solutions in similar situations (according to whether mentalistic or behaviouristic norms are applied).

I shall argue that this both alternatives are to be rejected, since a third, more appealing option is available: cognitive concepts can be interpreted in a flexible and neutral way, so that they are applicable also to some artificial entities. This would allow us to preserve both the spirituality and the unity of the law, even in a society increasingly characterised by automated information processing.

2 Three objections

Before developing the project of characterising cognitive states of SAs and analysing their legal implications, it is opportune to clear the field from three possible objections to this project.

Footnote 1 continued
companies dealing with intelligent agents, see http://www.aima.cs.berkeley.edu/ai.htmlagent. For an approach to artificial intelligence based on the idea of an agent, see Russell and Norvig (2003). For roadmap on agent.-based technologies (though limited to year 2005) see Luck et al. (2005).

The first objection is the following: from a legal perspective it is irrelevant whether people ascribe cognitive states to SAs, since the law adopts a behaviouristic perspective with regard to every agent (human or artificial). According to this view, the law would necessarily be behaviouristic, since cognitive states cannot be perceived, and therefore cannot be objectively ascertained by an impartial observer. The fact that the judge (and, more generally, any third party) cannot have direct access to another's minds, would imply that cognitive states cannot be legally relevant, i.e., that they cannot contribute to producing legal effects. It seems to me that this reasoning is based upon a fallacious inference. Each one of us (even the professional psychologist) has direct access only to the behaviour of other people, but this does not imply that we can never establish whether other people have certain cognitive states: we are often able to make this assessment on the basis of people's behaviour. The difference between a behaviouristic and a mentalistic approach does non concern accepting different cognitive inputs (behavioural inputs rather then mental inputs): the difference consists in that the first approach only registers people's behaviour, while the latter attributes cognitive states on the basis of behaviour. The latter approach views behaviour (having bought a gun) as providing clues for the presence of corresponding cognitive 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 operative facts, the judge and the lawyer cannot limit themselves to register observable facts (e.g., that the behaviour of a person caused the death of another), but they need to consider whether observable facts provide sufficient clues to establish cognitive attitudes (the intention to kill).[2]

The second objection is the following: attributing cognitive states to artificial systems implies unduly equating humans to such systems. According to this view, by attributing significance to the cognitive states of artificial systems we would implicitly deny that only humans have interests deserving legal protection, that only humanity is an "end in itself" (Kant 1996). I think that also this objection must be rejected: attributing legal relevance to the cognitive states of artificial entities does not imply attributing normative positions to such entities, in order to protect their own interests. In fact, here I shall only focus on whether the cognitive states of an artefact can contribute to determine legal effects on the head of natural or legal persons. I shall not address the completely different issue of establishing whether certain (future) kinds of artificial being may deserve legal protection on the basis of

[2] The thesis that the law gives some relevance to cognitive states does not entail that such states are always decisive: the law often needs to take into account, besides the perspective of the author of the act, also the way in which the act is understood by the counterpart and by third parties, as well as various pragmatic constraints. For instance. the law of contracts gives some limited relevance to the cognitive states of the author of a contractual declaration (by requiring that the party should intentionally make such declaration, in the awareness of its effects, and by allowing the contract to be annulled when certain mistakes were made in coming to a determination or in expressing it) even though such relevance may be overridden by further considerations, such as protecting the reliance or the counterpart (who justifiably assumed that the contract was regularly formed), or reducing litigation and facilitating the work of the judges (which may suggest that only in exceptional circumstances judges should override the text of the contract and its conventional linguistic meaning). The latter view is often said to characterise the British tradition, see Devlin (1962).

their intrinsic value, i.e., as entities deserving moral respect (see Taddei Elmi 1990; Karnow 1994; Solum 1992, 1255ff; Chopra and White 2004).

The third objection is the following: legal doctrine should not analyse cognitive concepts since this presupposes adopting a complete and uncontroversial theory of mind and conscience, which can substitute the intuitions of common sense. By engaging in this analysis a jurist would betray his/her mission, which is not that of making (or endorsing) controversial scientific hypotheses, but rather that of elaborating proposals which may be successful in legal practice, i.e., that of providing lawyers with the chance of converging into reasonable shared models. And one cannot expect lawyers to converge into adopting one theory of mind, i.e., to succeed where philosophers, psychologists and neurologists have so far failed to converge. It seems indeed that when one submits mental concepts to philosophical and scientific examination of mind and consciousness,[3] the certainties of common sense (which are the starting point for the lawyer) dissolve, and are substituted by complex and conflicting theories. My reply to this objection is that here I am not trying to defend any "deep" theory of mind to the exclusion of other theories, but more modestly, to articulate some rather superficial considerations concerning agency and cognitive attitudes, while respecting common sense.[4] For this purpose, I shall adopt the framework of Dennett (1997, 28 ff), and I shall distinguish three possible stances, the physical stance, the design stance and the intentional stance and consider what legal conclusions this may suggests with regard to artificial agency.

3 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.

For instance, we may explain the behaviour of a falling object on the basis of our knowledge of the physical laws of motion. Assume that I, emulating Galileo Galilei, throw a stone from Pisa's leaning tower: knowing that the acceleration of gravity is about 10 m/s, I can conclude that the stone that will have, after 0.5 s, the speed of about 5 m/s. In the same way, I can explain the shock 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 roadbed.

The physical stance can be applied not only to inanimate natural objects, but also to artefacts, animals or humans. Assume that unfortunately I let my mobile telephone drop from the top of the leaning tower, or that I myself am even more

[3] For a collection of some important contributions to the philosophy of mind, cf. Cummins and Dellarosa-Cummins (2000); for a basic introduction, cf. Davies (1998).

[4] See Peczenik (2006, 79), according to whom legal justification should be as much as possible "philosophically neutral" and jurists should "avoid commitment to strong philosophical theories and prefer weak philosophical theories."

unfortunately pushed into the air: I can provide an explanation (of forecast) of the speed of these falling objects according to the same physical laws. In fact, every object (be it natural, artificial, mechanical, electronic or biological, living or inanimate) obeys physical laws, and its behaviour can in principle be explained according to them.

Such explanations, however, though possible in theory, are practically feasible only to the extent to 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 situation of the bodies I am considering, I will not be able to make a forecast from the physical stance.

4 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 the piezoelectric lighter i just bought will emit a sparkle when I push the button on its back? Certainly not because I know the internal structure of my lighter and the physical laws governing it, but rather 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 artefact (a toaster, an automobile, a computer, etc.) from the design stance, I assume that the artefact, having been designed to perform certain functions will really work in such a way as to achieve these functions (if used in the way 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 artefact and therefore contribute, according to such functions, to the working of the artefact, in the way intended by its 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 distinguishing 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 includes two components: the control unity, 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 executed by their subcomponents, like, for the arithmetical-logical unit, the registers storing instructions, operands and results, the circuits carrying out the various instructions executed by the processor. 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 whose knowledge of mathematics and physics suffices to explain the functioning of all hardware components. Moreover, even these people cannot possibly know all conditions relevant to applying all physical laws (the electromagnetic condition of every single component of the computer).

In conclusion, when we are interested in the macro-behaviour of a complex artefact, 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 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 realising a particular operation, such as floating point multiplication, could have been wrongly designed so that, for a particular combination of input numbers, they provide an incorrect result. An artefact 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 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: 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 certain 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 connections activating 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, but the latter 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 available to us before malfunctioning has taken place.

The functional analysis characterising the design stance is applicable not only to artificial entities (which are built according to the design of their creators), but also to biological ones. 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.[5]

Firstly, only the organisms that 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 variations favouring survival and reproduction tend to be transmitted to a higher number of successors. Consequently the world would tend to be populated by organisms 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 would not 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 a function, which, according to some authors 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 1995, 13)—can also be generalised in such a way to be independent from such a reference.[6] Besides being applicable to

[5] I apologise for this trivialisation of the complex problem of evolution. For a "philosophical" introduction to Darwin's ideas, see Dawkins (1989) and Dennett (1996).

[6] For example, Nozick (1993) introduces the notion of a function on the basis of the concept of an homeostatic system, which he characterises 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." As examples of homeostasis, consider how an increase of bodily temperature may lead to sweating, which lowers the temperature, or 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

artefacts and biological organisms, the design stance can also be applied to social organisation, both in its purpose-based version 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 are replicated.

For instance, I may 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.

The design stance—through allowing us to make explanations that are normally correct (usually an artefact 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 have taken place. Moreover, even when an entity would work appropriately in its original environment, it may malfunction in a modified environment.

5 The intentional stance

Let us now move to the third and most controversial perspective, that is the intentional stance. Note that here the term "intentional" is used in the technical meaning it has in philosophical language, where this term typically refers to the relationship (also called "aboutness") between cognitive states or linguistic objects, and the things to which they refer to. 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 1987).

When looking at an entity from the intentional stance, we are explaining the behaviour of that entity assuming that it has certain cognitive states, 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) on the basis of certain representations of its environment

Footnote 6 continued
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 build 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 produced and maintained by natural evolution, which tends to make so that lungs are able to absorb oxygen (through the survival, and therefore reproduction, of the individuals whose lungs can absorb oxygen).

(beliefs). The behaviour of an intentional entity will then be explained as resulting from that entity attempting to achieve its goals, through means it believes to be appropriate. Further aspects of intentionality, which extend the belief-desire model, concern adopting intentions (in the specific sense of commitments to future action, see Bratman 1987) and norms.

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 (e.g., about what moves are available to its adversary), and assuming that it can devise rational ways to achieve these 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 1989, 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 the behaviour. of others. 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 these 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 the plant 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 describe the functions which are performed by the protein cover of that 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 provided by intentional interpretations.

Obviously, the intentional stance is the more useful the more the concerned entity is able to select, from a large range of possibilities, the behaviour 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 circumstances. 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 the reactions that 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 certain of its components, or certain of its behavioural reactions) with mutations, 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 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 may be correct to apply it towards living micro-organisms, at least as a first approximation.

The intentional stance is certainly appropriate with regard to certain complex artefacts, and in particular, with regard to certain computer systems. To use the classical example by Dennett, let us consider again computer systems for chess playing. These software programs 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 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, whenever she may do that in such a way that her losses are inferior to mine (and the operation has no negative side effects for her). 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 a convenient opportunity for eliminating some of my pieces, 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 then a human being? Shall I adopt the same strategy for interpretation and prediction (the intentional stance), and therefore attribute intentions and beliefs, 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 that would be 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 trying 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 that may be winning over the strategy that I assume Deep Blue is following.

Similar considerations also apply 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 cognitive attitudes (beliefs, intentions and possibly even emotions) and interpreting correspondingly their behaviour.

6 The intentionality of organisations and mixed entities

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 decision 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 these intentions and expectations, and even if I believe that such attitudes cannot be fully attributed to any individuals. 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 should reject both the theories that view the subjectivity of organisations as a mere fiction and the theories that view it as resulting from purely legal mechanisms: it may rather be based upon the need to adopt the intentional stance to explain and predict the behaviour of entities able to know, choose and act.

The entity viewed 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 performing different tasks: a software interacts with customers (drafts and sends sale offers, receives and confirms acceptation 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 parameterise 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 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' tastes to the address of one individual customer). The system as a whole appears as a unit of agency (for example, to people accessing its web site) and may attributed 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 ascribed to any isolated element of it.[7]

[7] On idea that combinations of humans and artificial entities may represent a new kind of hybrid subjectivity see Teubner (2006), referring to Latour (2005).

7 Intentional stance and SAs

The intentional stance is usually the only perspective from which we may hope to understand and forecast what an 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 an SA will normally have little knowledge of these 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 appropriate to the purposes assigned to it, on the basis of the information available to it, in the context in which it is going to operate, that is, such an interpretation will be based upon 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 an SA, *a fortiori* it needs to be adopted by the SA's counterparts in exchanges (where the SA is acting on behalf of its user/owner). The counterpart 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 at run 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 an 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 the SAs, in order to achieve the objectives assigned to them, and by using the knowledge they have, will get to the determinations they declare to their counterparts, according to existing linguistic and social conventions. So, the assumption of rationality (relative to the cognitive states of the SA) still provides the default background for understanding the SA's behaviour.

8 The nature of intentional states

Before concluding the discussion of the intentional stance, let us approach the difficult issue of determining the ontological status of intentional attitudes: what is the reality described by asserts attributing intentional attitudes (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, on which I can only make some very superficial 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 that entity (and postulating that the entity can behave rationally on the basis of these 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 more favourable position, considering possible replies of the adversary. The second consists of a large table, that connects every possible situation in the board to a specific move. Since the two systems behave exactly in the same way, a behaviourist approach 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 specific internal features of the entity to which they are attributed. Consequently, to establish whether an entity truly possesses cognitive states (goals, beliefs, intentions, etc.) one needs to consider whether there are specific states 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 functioning.

Obviously, a realistic attribution of intentional states to hardware or software artefacts or to organisational structures, assumes that we can identify appropriate 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 these situations (in such situations 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 these 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 these situations, and
- this covariance enables the entity to react appropriately to the presence of these situations.

I cannot approach here this difficult issue (on covariance, cf. Dretske 1986, and for a discussion of the literature, cf. Davies 1998, 287 ff). Let me 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 the situations in which 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 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 one types 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, the 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 that 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 these 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 adequate nourishment. Similarly, assume that an SA has been given the description of certain goods and that, on the basis of such a description, it starts operating in order to buy these goods, until the goods are purchased (which will lead it to remove the description from its task queue). Under such conditions, we can plausibly 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 an 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 results, 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 counterparts 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 counterpart assumes it believes, and has it has the goals which the counterpart assumes it has?).

9 Intentionality, consciousness and normativity

Some authors link the notion of intentionality to the notion of consciousness or awareness. For example Searle (1989, 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 1990, 586). I think that this approach should be rejected, since it forces us to renounce to the intentional stance in regard to all non-human entities (unless we want to trivialise the notion of consciousness). To give an intentional interpretation to the behaviour of such entities we must conceptualise our cognitive notions in such a way that they are also applicable to certain artificial systems. This does not exclude that other, richer notions (such as the notion of consciousness) remain applicable only to human beings (and possibly human organisations).

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 1997, 119 ff). Such a capacity can be fully attributed, besides than to humans, also to human organisations. On the other hand the intentional stance remains applicable also to entities (such as animals and SAs), to which we cannot attribute reflexivity.

In the same way, we need to reject a necessary connection between intentionality and normativity, i.e., the idea that attributing intentionality to an entity presupposes that the entity has the capacity of following norms or rules, even when they clash against its desires. The intentional perspective can 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 considerations I have developed above suggest positive conclusions concerning the possibility of attributing to artificial entities the specific type of intentionality which underlies the execution of speech acts, and in particular declarations (of will or intention), like those performed by an SA proposing or accepting to purchase or sell a certain item. This attribution only presupposes that it is possible to explain the agent's behaviour (suggest the performance of an utterance), assuming that the agent intends to produce certain normative results through uttering a certain statement, believing that this utterance (accompanied by such intention) will produce such results.[8]

10 The intentional stance and the law

In the previous pages I 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

[8] For interesting considerations on the attribution of agency and intentionality to artificial entities, an attribution which—contrary to the approach here taken—is based on communicative capacity rather than on purposive rationality, and more on socially shared presumptions than on the nature of the concerned entities, see Teubner (2006).

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 were 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 (and having scientific knowledge, not available today, on the connection between brain states and behavioural propensities).

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 expectations (reliance or trust) that people have as a consequence of adopting such stances (on trust, see Jones 2002; Castelfranchi and Falcone 2005).

Let us first consider the design stance. When we look at an artefact that seems to embody a certain design, why do we expect it to work according to that design? Why are we ready to take the risk that the behaviour of the artefact does not correspond to such design? Different factors converge in supporting one's expectations that the behaviour of the object corresponds to its assumed design, but among these factors there are also some legal rules. For example, the law requires that the seller should guarantee that the thing has no faults; it states that the producer is liable for damages caused by malfunctioning; it requires that the owner or guardian of a thing should be liable for damages caused by its anomalous behaviour. The legal protection of expectations grounded on the design stance is realised through putting the obligation to compensate damages upon the subject who could ensure the satisfaction of these expectations. 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 I 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 artefacts with which we need to interact.

Let us now consider the expectations we form when looking at social reality from the intentional stance. Why should we interpret other people's expectations by ascribing beliefs, desires and intentions? Why should we take the risk that interpretations and forecasts based upon the intentional stance are disappointed, that the persons to which we attribute a certain cognitive state do not behave according to the expectations grounded upon that 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 author having certain cognitive states (usually intentions and beliefs): the intention to make a certain declaration, the intention to realise the declared normative states, (true or false) beliefs concerning the relevant facts. On the other hand, the effects of an act are also conditioned to the intentional states that the counterpart ascribes to the author, of the basis of the available clues, as provided by the act itself (and in particular by its linguistic content), but also sometimes by the act's circumstances, including the previous behaviour of the author. If there is a difference between the cognitive states possessed by the author of an act and the cognitive states attributed to him/her by the counterpart, the decisive criterion can be represented by social conventions, which define the meaning that one party can legitimately ascribe to the behaviour of the counterpart.

Such triple intentional qualification of an act (the point of view of the author, of the counterpart, of the existing social conventions) leads to conflicts which are well known to the students of the formation of contracts. The law usually decides such conflict by focusing on the protection of reliance (trust): one party's erroneous belief that the counterpart made a certain declaration (which the counterpart did not make) or had a certain intention (which the counterpart did not have) often produces the same effect that the existence of the counterpart's declaration or intention would have produced. This happens to protect the party who was justified in believing that the counterpart made that declaration or had that intention, which is usually the case when this belief corresponds to existing socio-linguistic 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. I 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 (and consequently act) 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), obtained by interpreting according to social conventions the clues she has provided to me, since I know that, even if my good-faith ascription were wrong, there would be still the legal effects which would have taken take place in case the ascription were true (if my partner really has such intentions and beliefs). 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 needed to prevent possible misunderstandings, and I therefore can assume that she really has the cognitive states that she appears to have.

Also other legal rules tend to ensure one's capacity to attribute intentional states to one's partners in various social interactions. Consider for example legal rules punishing the malicious communication of false information: here the law protects the reliance in other people's assertions, transferring the costs 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, the intentions of a contractor 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 a meaning to one clause and at the same time believe that my counterpart attributes a different meaning to the same clause. Similarly, in mistake and deceit, the law attributes relevance to one party's knowledge that the counterpart does not know a certain fact (if the contractor knows that the counterpart is mistaken in regard to an essential term of the contract the contract may be voidable).

11 The intentionality of artefacts

The intentional stance represents usually the only possible viewpoint to explain and foresee the behaviour of complex entities that can act teleologically. Consequently, the law should not refuse to acknowledge the intentionality of artefacts having this capacity. The recognition of the intentionality of computer systems (and of hybrid systems, including both humans and computers) would imply two sets of consequences:

On the one hand the counterpart interacting with such a computer system would be authorised to attribute to it the intentional states that it appears to have, and in particular, the intentional states that the system declares to possess or that are presupposed by the speech acts it accomplishes. According to the principle of protection of reliance, the owner of the system will not be able to avoid that the system is assumed to have intentional states that (a) have been attributed, in good faith, to the system by the counterpart 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 an SA performs a speech act that 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 chooses 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 tort liability). Similarly, if an 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 counterpart of a computer system will not be able to reject 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 an 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. Moreover, 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 SA. As this last example shows, the intentional stance can be adopted not only towards the

intentions of a computer system that are directly expressed in the declarations of the system, but also towards the cognitive states presupposed by such declarations. So, when an SA makes a contractual declaration (for example, it declares the willingness to buy 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 cognitive states that are presupposed by such an 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).

12 The delegation of cognitive tasks

The discussion of the issue of intentionality and cognition in computer systems has led to the conclusion that SAs (as other computer systems) can perform cognitive processes and can be attributed intentional states. Therefore, delegating them cognitive tasks seems the appropriate way of using them, and viewing them as having been delegated such tasks seems the appropriate way of interacting with them.

From this perspective the reason why the effects of what is done by an SA fall upon the user is not the fiction that the user has wanted or has foreseen the behaviour of his SA but rather the fact that the user has chosen to use the SA as a cognitive tool, and has manifested to possible counterparts this determination as well as the determination to bear the legal effects of declarations made by the SA on the user's behalf, on the basis of the SA's cognitive activity (cf. Sartor 2003). 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 counterparts), but this is the will to delegate certain cognitive tasks to the SA, tasks to be performed on the user's behalf. It is not the will of performing every specific action accomplished 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 counterparts, 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 not attributable to the user distinguishes SAs from other objects or tools, also from a legal perspective.

It is true that the phenomenon here discussed (cognitive delegation to an artefact) is not completely new: cognitive delegation takes place (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 user's mistake into account. This is not the case, however, when an SA is charged with accomplishing a legal activity directly, i.e., when "no natural person reviewed or

intervened in each of the individual actions carried out by the automated message systems or the resulting contract" (art. 12 of the "United Nations Convention on the Use of Electronic Communications in International Contracts Document", adopted by the General Assembly of the United Nations on 23 November 2005). Under such circumstances we need to address directly the issue of cognition performed by artefacts, and consider what its legal relevance may be, according to the nature of these artefacts, and the (reasonable) expectations of their users and interlocutors.

It is also true that 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 though Internet sites (the classical reference for Italian legal doctrine is Cicu 1901, addressing sales though vending machines). I am 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 just mentioned. In the following I shall consider the legal implication of the approach just sketched for different areas of the law, focusing especially on tortious liability, contracts and personality.

13 SAs and tortious liability

In order to provide compensation for damages caused by an SA a custodian must be identified, who may bear the obligation to restore. With regard to a material thing identifying a custodian is often easy: this is the thing's 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 consider that different subjects may "own" different aspects of an 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 correct approach at least when one adopts a proprietary approach to personal data, as suggested for example, by Lessig 1999, 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 I have always put the expression "owned" between inverted commas, to indicate that each one of these 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. 2001; Yip and Cunningham 2002). In any case, since the user has no control over certain aspects of his SA, it may be unfair to regard the user as a custodian, in relation to damages related to these 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 is even forbidden to decode and modify it. To approach liability for failures concerning such aspects, the usual idea of custody is insufficient: we need either to 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 to supplement it by appealing to different branches of the law (e.g. product liability, consumer protection, etc.).

With regard to the extent of the custodian's liability two approaches are possible. According to the first approach, 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 an 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 an 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 (as observed by Allen and Widdison 1996). 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, I have remained within the boundaries of well-known legal problems, where the issues related to SAs are not so different from those pertaining to other technological objects.

The "new" issues to be addressed concern the feature of SAs introduced above, i.e., the fact that they are cognitive tools. We need to consider whether the cognitive states of an SA are relevant to establishing and circumscribing the liabilities deriving from damage caused by that SA. Note that this does not amount to asking whether an SA is legally responsible, and even less to ask whether it is morally responsible. This is irrelevant, since by "liable" (legally, morally, or whatever) I just mean "obliged to pay compensation", and the only liability I am considering is the user's liability. I shall 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 "caused" by an SA, so that the user will be liable to compensate the ensuing damages (regardless of fault, if strict liability applies).

Consider for example the case of an 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? A criterion to limit causality is the idea of "normal" or "adequate" causality: an event only "causes" its normal effects: an exceptional effect, due to extraordinary concurring factors (such as the system's malfunctioning, in our example) does not really count as being caused by the event. However, the limitation of causality to "normal" effects leads to absurd results in the case of damages intentionally (deliberately) produced in exceptional circumstances, known to the author of the damaging act, even when the intention at stake is the SA's rather then 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 SAs, which consists in performing a normally innocuous 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 there are sufficient clues for ascertaining the existence of 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 innocuous (though damaging in the particular case and intentionally malicious) behaviour of his/her 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 (known to the SA), while damage unintentionally caused could 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 an 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 these duties. It only implies that if the SA does not behave as 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 caution), 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 be 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 that the intentions of the SA may condition the user's liability. If an SA deliberately caused damage, then this implies that this damage could have been cheaply avoided: the user could easily have constrained the behaviour of his system in such a way that it would refrain from take such malicious initiatives.

So, it seems that the guardian's liability for the action of an 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 an SA would be similar, rather then 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, while acting in the course of the employment. Note that the relevance of the SA's cognitive states is the reason why one may assimilate the relation of an SA to its user to the relation of an employee to his employer. This has nothing to do with labelling an SA as a person, or as a legal or moral subject.

14 SAs and contracts

On line contracting (and bargaining) is already a very important application area for agent-based technologies.[9] This is confirmed by the fact that some legislatures have already shown some interest for this domain. The US Uniform Computer Information Transactions Act (UCITA), aimed at complementing the US Uniform commercial code with regard to software contracts, establishes some rules that specifically concern SAs. UCITA defines an agent as "a computer program or electronic or other automated means, used independently to initiate an action, or to

[9] On the legal aspects of contracts made by SAs, there is already a significant literature, see for instance Kerr (1999); Lerouge (2000); Bellia (2001); Weitzenboeck (2001); van Haentjens (2002); De Miglio et al. (2002); Weitzenboeck (2004); Wettig and Zehendner (2004); Kafeza et al. (2005); Barfield (2005); Andrade et al. (2007); Balke and Torsten (2008).

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."[10] The European E-Commerce Directive, does not explicitly mention software agents, though a reference to them was included in the explanatory notes accompanying the proposal of the directive, specifying that that Member States should not "prevent the use of electronic systems as intelligent electronic agents"). However, the directive implicitly supports the use of SAs in contracts, by requiring Member states to ensure that electronic contracts be not "deprived of legal effectiveness and validity on account of their having been made by electronic means." Similarly, the above mentioned UN Convention on Electronic Communications in International Contracts Document speaks of an "automated message system"to refer "essentially to a system for automatic negotiation and conclusion of contracts without involvement of a person, at least on one of the ends of the negotiation chain", affirming that no human review is required for the validity of such contracts.

With regard to contracts SAs charged with negotiating and concluding contracts in the name and in the interest of their users, I shall argue that we need to take seriously the idea that an SA may have cognitive states 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 so, since, as 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 an SA fall upon the user is not explained by the fact that the user foresaw the behaviour of its SA, or could have foreseen it, or even ought to have foreseen it (as affirmed, for example, by Finocchiaro 2002). It is true that the intention of the user (as recognised by the counterpart) provides the ultimate justification for the effectiveness of the contracts made by his SA, but this is his/her intention to entrust the SA with the task of entering into certain kinds of transactions in the user's name, performing the cognitive processes that are required for preparing and executing these 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 a particular contract is made, nor any knowledge of its specific

[10] UCITA has been very controversial (especially since it allows contracts to override consumer protection rules) and been adopted so far only by two States, while being rejected by the American Law Institute.

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 1996). 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 1998), 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 counterpart has certain cognitive states, should be irrelevant to the effects of these 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 an SA has formed a certain intention, or had certain beliefs, at least when this was known to the counterpart, may impinge on the effects of the contract.

Consider for example the following hypothetical. Assume that an SA has been charged with the task of selling on line certain pieces of old jewellery according to their weight, age, and material. 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 counterpart accepts. Assume also that in the photograph of the ring available on line one could easily see the words "solid 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 counterpart, or will the contract be valid, since an SA cannot have any cognitive states, and therefore cannot make any mistake? In general, legal systems allow similar contracts to be voided when a human mistake was important and was recognisable to the counterpart. What will happen when, as in the case at hand, the mistake was committed by an SA? And what if the counterpart knew that the SA was making a mistake? And what if the SA's mistake was induced by the counterpart, 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 considered above, since appropriate offer and acceptance messages were sent, the behaviourist conclusion will be 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 we cannot 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 ensue from having certain cognitive states, then these effects will sometimes be produced even when the corresponding cognitive states are absent, and they may not be produced 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.

We should rather take the intentional stance, and attribute cognitive states to SAs, considering whether having these states (and being attributed them) should make a difference in the legal effects of the behaviour of such systems. Thus, the counterpart of an 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 an SA makes a certain declaration, in appropriate circumstances, would be a sufficient clue to the SA's intentional states. Moreover, whenever the counterpart 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 counterpart 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 an SA, as impairment to the formation of the contractual volition, should have legal implications similar to the so-called defects of will (mistake, deceit, duress).

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 counterpart) 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 these rights and obligations, but because he has chosen to delegate to his SA the formation of contacts 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 an 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 1988 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 its user: when it is necessary to establish who wanted what, we need to examine which contents of the contract where pre-established by the user, and which ones where 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 must 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 counterpart. 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 an SA. If an SA's mistake is not recognisable to the counterpart, 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. This view seems compatible with the following statement, included in the Unicitral document accompanying the proposal of the UN Convention Electronic Communications in International Contracts 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, par. 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.

SAs are indeed no sale agents strictly understood, and we cannot automatically transfer to users of SAs every rule applicable to principals of human sales agents (since in particular, sales agents may be liable on their own). However, both a sale agent and an SA have been delegated a cognitive task, and this may justify coming to the same legal conclusions, to certain regards. Assume, for instance, that a mobile SA moves into a financial marketplace, and then proceeds to buy and sell stock, without interacting with its user and with the computer system of the latter. If the SA were only a means for the user to communicate with other parties, contracts 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 would preclude the mobile SA from selling what it has just bought, before communicating the purchase to the user (since communication to the user is necessary to perfect the previous exchange), and therefore would preclude engaging in effective on-line trading. In this case, the model of the sale agent acting as a representative provides the right clue: for finalising the contract it should be sufficient that the counterpart's acceptance reaches the representative (the SA).

15 SAs and personality

To address the attribution of personality, we need to distinguish three normative positions:

1. the ability to have one's own legal positions, i.e., the ability to have rights and duties of one's own;
2. the ability to produce, through one's intentional actions, rights and obligations on one's head;
3. the ability to produce, though one's intentional actions, rights and obligations on the head of another.

Only the first two positions characterise legal personality, broadly understood. The third one (which we have examined in the previous sections) is independent from the others: having legal personality does not entail that one is able to bind another; this usually presupposes a delegation by the concerned person (or a different specific legal ground).

Thus, giving an SA legal personality means that it has the ability to have rights and duties of its own, and the ability to acquire or transfer these rights and duties through contracts (items 1 and 2 above). If an SA were considered a legal person, then it would be able to enter into contracts in its own name and to acquire its own rights and duties, which it might later transfer to its user or to a third party (consider for example an SA acting as an on-line trader, buying certain commodities and reselling them at a higher price). 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, by transferring 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 counterparts, 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 counterparts would not know on whose behalf the SA is acting. If the SA does not fulfil its obligations, we may then imagine that the SA's creditors 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 a complete separation of the patrimony of the SA from the patrimony of its users. If an 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 an SA? Secondly, SAs may disappear, definitely (being cancelled) or temporarily (being registered on an inaccessible storage device), they can divide themselves into the modules they include, 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 on its obligations).

Some of these problems can possibly be solved through some legal artifices: for example the residence or the domicile of an SA may be the address of the bank where the SA's fund is deposited, the SA would be attributed the acts 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 at present 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 the SAs. The counterparts of an SA could then be warranted by the capital of the company and by the legal remedies available against defaulting commercial companies.

Finally, 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 (on the basis of the contract between the user and the marketplace), if the SA violates the rules of the marketplace. Thus, regardless of whether SAs can be viewed as the addressees of legal norms (though not having rights and duties on their own), it makes sense to speak of normative SAs, and design SAs having the cognitive competence for adopting and complying with norms (cf. Castelfranchi et al. 1999; Artosi 2002; Boella and Damiano 2002; Brazier et al. 2002; Gelati et al. 2002).

16 Some legal issues concerning SAs

Contracts, torts and personality do not exhaust the legal issues concerning SAs. In the following pages I shall concisely consider some further more specific topics that have attracted the attention of legal doctrine:[11]

- SAs and consumer protection,
- SAs and intellectual property,
- SAs and privacy protection,
- SAs and right to information.

16.1 SAs and consumer protection

In discussing SAs and consumer protection, I will refer to Italian law, which is similar to the law of other developed countries (see, for example van Haentjens 2002; Rossato 2002). 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 counterpart). Moreover, certain contractual terms, which are likely to impair the position of the customer, are ineffective in contracts between a professional operator and a consumer (according to art. 1469 bis of the Italian civil code, which implements European legislation). 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 the customer's SA) could negotiate with the seller's SA, and accept a change in 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 can be practically costless) would make irrelevant every law establishing contractual terms which can be derogated by the parties, since these 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 idea of Coase (1960), who argued that with no transaction costs, economic efficiency alone decides the allocation and the use of resources.

[11] Further significant legal issues concern the SAs' use in virtual enterprises (cf. Cevenini 2002), in on-line dispute resolution (cf. Chiti and Peruginelli 2002; Gouimenou 2002), or in police investigations (Burkhard 2006; Abel 2009).

One further aspect of consumer protection concerns whether software agents can engender specific risks for consumers. In particular the SA of a merchant could simulate familiarity and companionship, and thus induce an appearance of friendship that can facilitate the exploitation of the consumer (Kerr 2004).

16.2 SAs and intellectual property

With 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 (for a discussion of this topic, cf. Bing and Sartor 2002, 2003). 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 these 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 an 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 author-SA, 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?

16.3 SAs and privacy

In the privacy domain (cf. Borking et al. 1999; Bygrave 2001; Villecco 2002; Brazier et al. 2004; Boonk and Lodder 2006), 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, an 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, the knowledge of which may provide an unfair advantage to the counterpart, even when the SA's 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 willing to pay).

16.4 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 1999, 164 ff; Sunstein 2001). 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 should be made explicit to their users (in an understandable form) and that there should be 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.

17 Conclusion

The subject of SAs suggests bold speculations on futuristic scenarios. For instance, Kurzweil (1999) forecasts that soon the distinction between humans and SAs will become 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 will be able to move from a biological to an electronic substrate and vice versa? What will be the legal relationship between the various embodiments of one individual?

Here I shall abstain from such science-fictional conjectures, but I shall nevertheless address a possible future conflict between SAs and human values, i.e., the fact that the widespread use of SAs might diminish our human and social competence. Once we admit SAs in typically human relations, we might 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 might contribute to compromising the chance of establishing authentic relationships with other persons. Shall the Hegelian dialectic between master and slave (as presented in Hegel 1931, par. 189ff.) be reproduced with regard to the relation between human and their artificial agents (namely, digital and physical robots)? Shall we delegate so much to them, and become so dependent on them that we will lose our ability to think and act on our own? Shall we so much interpose our electronic slaves between ourselves and the satisfaction our desires (as Hegel would say), that we become completely passive, mere "desire machines", having transferred to such slaves all productive and communicative initiatives required for satisfying such desires.[12]

It seems to me that these worries can be countered by observing that the substitution of intelligent machine for humans in creative tasks is very far away: a symbiotic cooperation (involving a considerable amount of cognitive delegation) between humans and machines is rather to be expected in the near future (on the symbiosis between humans and machines see Licklider 1960). Moreover, the SA model does not identify a set of specific software products, it is rather a comprehensive approach to computing, a paradigm which may lead to very different applications. Thus, we need to ensure that the use of this paradigm will provide us with trusted electronic helpers without forcing us to renounce our capacity for decision and interaction, and that it will increase rather then diminish security and trust. The realisation of these objectives also depends on the definition of an appropriate legal framework. This framework, however cannot be designed in the abstract. It needs experimentation with problems and solution, which presupposes both a liberal approach to the use of SAs in legal activities and the ability to react with appropriate legal remedies when users' interests and legal values are endangered.

References

Abel W (2009) Agents, trojans and tags: the next generation of investigators. Int Rev Law Comput Technol 23:99–108

Allen T, Widdison R (1996) Can computers make contracts. Harvard J Law Technol 9:25–52

Andrade F, Novais P, Machado J, Neves J (2007) Intelligent contracting: software agents, corporate bodies and virtual organisations. In: Camarinha Matos L, Afsarmanesh H, Novais P, Analide C (eds) Establishing the foundations of collaborative networks. Springer

Artosi A (2002) On the notion of an empowered agent. In: Proceedings of LEA 2002. Workshop on the law of electronic agents. Bologna, CIRSFID, pp 123–131

[12] The idea that relying on automated helpers may compromise human values is at the centre of the work of Isaac Asimov (see for instance, Asimov 1996). For an optimistic view concerning the impact of robots on human society, see Brooks (2002).

Asimov I (1996) Foundation and earth, 1st edn. Harper-Collins, London (1986)
Balke T, Torsten E (2008). The conclusion of contracts by software agents in the eyes of the law. In: Padgham L, Parkes DC, Müller J, Simon P (eds) Proceedings of 7th international joint conference on autonomous agents and multiagent systems (AAMAS 2008), vol 2. IFAAMAS, pp 771–778
Barfield W (2005) Issues of law for software agents within virtual environments. Presence Teleoper Virtual Environ 14:741–748
Bellia AJ (2001) Contracting with electronic agents. Emory Law J 50:1047–1092
Bing J, Sartor G (2002) Lovely Rita: a scenario. Deliverable, ALFEBIITE (IST-1999-10298)
Bing J, Sartor G (eds) (2003) The law of electronic agents. Oslo, Unipubskriftserier
Boella G, Damiano R (2002) 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. Bologna, CIRSFID, pp 111–121
Boonk ML, Lodder AR (2006) Halt, who goes there? On agents and conditional access to websites. In: Proceedings of BILETA 2006, globalisation and harmonisation in technology law
Borking JJ, van Eck BMA, Siepel P (1999) Intelligent software agents and privacy. The Hague, Registratiekamer
Borruso R (1988) Computer e diritto: Problemi giuridici dell'informatica, vol 2. Milano, Giuffrè
Bratman M (1987) Intentions, plans and practical reasoning. Harvard University Press, Cambridge
Brazier F, Kubbe O, Oskamp A, Wijngaards N (2002) Are law abinding agents realistic? In: Proceedings of LEA 2002. Workshop on the law of electronic agents. Bologna, CIRSFID, pp 151–157
Brazier F, Oskamp A, Prins C, Schellekens M, Wijngaards N (2004) Anonymity and software agents: an interdisciplinary challenge. Artif Intell Law 12:137–157
Brooks RA (2002) Robot: the future of flesh and machines. Penguin, London
Burkhard S (2006) The taming of the sleuth-problems and potential of autonomous agents in crime investigation and prosecution. Int Rev Law Comput Technol 20:63–76
Bygrave LA (2001) Electronic agents and privacy: a cyberspace odyssey 2001. Int J Law Inf Technol 9:275–294
Castelfranchi C, Dignum F, Catholijn MJ, Treur J (1999) Deliberative normative agents: principles and architecture. In: Proceedings of ATAL 1999, pp 364–378
Castelfranchi C, Falcone R (2005) Socio-cognitive theory of trust. In: Pitt J (ed) Open agent societies: normative specifications in multi-agent systems. Wiley, London
Cevenini C (2002) Agents in the virtual enterprise: some legal notes. In: Proceedings of LEA 2002. Workshop on the law of electronic agents. Bologna, CIRSFID, pp 59–64
Chiti G, Peruginelli G (2002) Artificial intelligence in alternative dispute resolution. In: Proceedings of LEA 2002. Workshop on the law of electronic agents. Bologna, CIRSFID, pp 97–104
Chopra S, White L (2004) Artificial agents—personhood in law and philosophy. In: Proceedings of ECAI 2004. Amsterdam, IOS
Cicu A (1901) Gli automi nel diritto privato. Il Filangeri 8:1–30
Coase R (1960) The problem of social cost. J Law Econ 3:1–44
Cummins R, Dellarosa-Cummins D (eds) (2000) Minds, brains, and computers: the foundations of cognitive science. Blackwell, London
Davies M (1998) The philosophy of mind. In: Graylin AC (ed) Philosophy 1: a guide through the subject. Oxford University Press, Oxford, pp 250–335
Davis JR (1998) On self-enforcing contracts, the right to hack, and wilfully ignorant agents. Berkley Technol Law J 1148
Dawkins R (1989) The selfish gene, 2nd edn. Oxford University Press, Oxford
De Miglio F, Onida T, Romano F, Santoro S (2002) Electronic agents and the law of agency. In: Proceedings of LEA 2002. Workshop on the law of electronic agents. Bologna, CIRSFID, pp 23–32
Dennett DC (1989) The intentional stance. MIT Press, Cambridge
Dennett DC (1996) Darwin's dangerous idea: evolution and the meanings of life, 1st edn. Penguin, London (1995)
Dennett DC (1997) Kinds of minds: towards an understanding of consciousness. Basic, New York
Dennett DC, Haugeland JC (1987) Intentionality. In: Gregory RL (ed) The Oxford companion to the mind. Oxford University Press, Oxford, pp 383–386
Devlin P (1962) The principles of construction of charterparties, bills of lading and marine policies. In: Samples of lawmaking. Oxford University Press, Oxford, pp 52–66
Dretske F (1986) Misrepresentation. In: Bogdan RJ (ed) Belief: form, content and function. Oxford University Press, Oxford, pp 17–36

Finocchiaro G (2002) 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. Bologna, CIRSFID, pp 75–80
Gelati J, Rotolo A, Sartor G (2002) Normative autonomy and normative co-ordination: declarative power, representation, and mandate. In: Proceedings of LEA 2002. Workshop on the law of electronic agents. Bologna, CIRSFID, pp 133–150
Gouimenou J (2002) E-arbitration-t ©: an alternative dispute resolution for SMEs. In: Proceedings of LEA 2002. Workshop on the law of electronic agents. Bologna, CIRSFID, pp 105–110
Hegel GWF (1931) The phenomenology of mind. Allen and Unwin, London
Irti N (1998) Scambi senza accordo. Rivista trimestrale di diritto e procedura civile 347–364
Jones AJ (2002) On the concept of trust. Decis Support Syst 33:225–232
Kafeza I, Kafeza E, Chiu DKW (2005) Legal issues in agents for electronic contracting. Proceedings of the 38th Hawaii international conference on system
Kant I (1996) Groundwork of the metaphysics of morals. In: Gregor MJ (ed) Practical philosophy. Cambridge University Press, Cambridge, pp 37–107
Karnow C (1994) The encrypted self: fleshing out the rights of electronic personalities. John Marshall J Comp Inf Law 13:1–16
Kerr IR (1999) Spirits in the material world: intelligent agents as intermediaries in electronic commerce. Dalhousie Law J 22:189–249
Kerr IR (2004) Bots, babes and the californication of commerce. Univ Ottawa Law Technol J 1:284–324
Kurzweil R (1999) The age of spiritual machines. Orion, London
Latour B (2005) Reassembling the social: an introduction to actor-network-theory. Oxford University Press, Oxford
Lerouge JF (2000) The use of electronic agents questioned under contractual law: suggested solutions on a european and american level. John Marshall J Comput Inf Law 18 (2):430–500
Lessig L (1999) Code and other laws of cyberspace. Basic, New York
Licklider JCR (1960) Man-computer symbiosis. IRE transactions on human factors in electronics HFE-1 (March), 4–11
Luck M, McBurney P, Shehory O, Willmott S (2005) Agent technology: computing as interaction (a roadmap for agent based computing).AgentLink. (http://www.agentlink.org/roadmap/index.html)
Nozick R (1993) The nature of rationality. Princeton University Press, Princeton
Peczenik A (2006) Scientia juris. In: Treatise of legal philosophy and general jurisprudence, vol 2. Springer, Berlin
Pitt J, Mamdani A, Charlton P (2001) The open agent society and its enemies: a position statement and a programme of research. Telemat Inform 18(1):67–87
Rossato A (2002) "Stop the bot!": trespass to chattels in cyberspace. In: Proceedings of LEA 2002. Workshop on the law of electronic agents. Bologna, CIRSFID, pp 159–172
Russell SJ, Norvig P (2003) Artificial intelligence a modern approach, 2nd edn. Prentice Hall, Englewood Cliffs
Sartor G (2003) Gli agenti software e la disciplina giuridica degli strumenti cognitivi. Diritto dell'informazione e dell'informatica 27–59
Searle JR (1989) Consciousness, unconsciousness, intentionality. Philos Top 17:193–209
Searle JR (1990) Consciousness, explanatory inversion and cognitive science. Behav Brain Sci 13:585–596
Searle JR (1995) The construction of social reality. New York, Free
Solum LB (1992) Legal personhood for artificial intelligence. North Carolina Law Rev 70:1231–1287
Sunstein CR (2001) Republic com. Princeton, Princeton University Press
Taddei Elmi GC (1990) I diritti dell'intelligenza artificiale tra soggettività e valore: fantadiritto o jus condendum. In: Lombardi Vallauri L (ed) Il meritevole di tutela. Milano, Giuffrè, pp 685–711
Teubner G (2006) Rights of non-humans? Electronic agents and animals as new actors in politics and law. J Law Soc 33:497–521
van Haentjens O (2002) Shopping agents and their legal implications regarding Austrian law. In: Proceedings of LEA 2002. Workshop on the law of electronic agents. Bologna, CIRSFID, pp 81–96
Villecco A (2002) Agent technology and on-line data protection. In: Proceedings of LEA 2002. Workshop on the law of electronic agents. Bologna, CIRSFID, pp 53–58
Weitzenboeck E (2001) Electronic agents and the formation of contracts. Int J Law Inform Technol 9(3): 204–234
Weitzenboeck E (2004) Good faith and fair dealing in contracts formed and performed by electronic agents. Artif Intell Law 12:81–110

Wettig S, Zehendner E (2004) A legal analysis of human and electronic agents. Artif Intell Law 12:111–135
Yip A, Cunningham J (2002) Some issues on agent ownership. In: Proceedings of LEA 2002. Workshop on the law of electronic agents. Bologna, CIRSFID, pp 13–22

Author Biography

Giovanni Sartor is professor of Legal informatics and Legal Theory at the European University Institute of Florence and at the University of Bologna. He obtained a PhD at the European University Institute (Florence), worked at the Court of Justice of the European Union (Luxembourg), was a researcher at the Italian National Council of Research (ITTIG, Florence), held the chair in Jurisprudence at Queen's University of Belfast (where he now is honorary professor), and was Marie-Curie professor at the European University of Florence. He is President of the International Association for Artificial Intelligence and Law. He has published widely in legal philosophy, computational logic, legislation technique, and computer law. Among his publications is: Corso di informatica giuridica (Giappichelli 2008), Legal Reasoning: A Cognitive Approach to the law (Springer: 2005), The Law of Electronic Agents (Oslo: Unipubskriftserier, 2003), Judicial Applications of Artificial Intelligence (Dordrecht: Kluwer, 1998), Logical Models of Legal Argumentation (Dordrecht: Kluwer, 1996), and Artificial Intelligence in Law (Oslo: Tano, 1993).