Attribution of Knowledge to Artificial Agents and their Principals*

Samir Chopra
Department of Computer and Information Science
Brooklyn College
Brooklyn, NY 11210
schopra@sci.brooklyn.cuny.edu

Laurence White
Rue Franklin 157
Brussels B-1000
Belgium
laurencefwhite@gmail.com

Abstract

We consider the problem of attribution of knowledge to artificial agents and their legal principals. When can we say that an artificial agent *X* knows *p* and that its principal can be attributed the knowledge of *p*? We offer a pragmatic analysis of knowledge attribution and apply it to the legal theory of artificial agents and their principals.

1 Introduction

An agent's principal is its employer, or any other legal person engaging the agent to carry on transactions on its behalf. A problem commonly faced by courts is deciding when to attribute the knowledge in possession of an agent to its principal. If the agent in question is an *artificial* one, how should the courts decide that a) the agent *knows* the proposition in question and b) that *this knowledge can be attributed to the agents' principal*? We need a philosophical account of knowledge attribution that does justice to the first question, and thereby aids in the resolution of the second – legal – problem. Conversely, legal resolutions of these issues will aid us in a solution of the philosophical problem – just as legal findings on personhood can clarify philosophical debates over the nature of personal identity. As the delegation of tasks to artificial agents increases, so will cases that encounter the need for decisions hinging on these debates, thus rendering more urgent the need for a solution.

The plan of this paper is as follows. In Section 2, we present our analysis of knowledge attribution. In Section 3, (within limitations of space) we illustrate its plausibility by examples and arguments. In Section 4 we describe the legal problem of attributing knowledge held by agents to their principals (while critically examining current legal doctrine in this area) and show how our analysis is of help. While we concentrate on British Commonwealth and US law, this paper does not purport to be a complete survey of the relevant law in the jurisdictions dealt with. When citing non-academic sources such as case reports, we use the legal mode of citation.

* We thank Rohit Parikh, John Sutton and the referees for helpful comments.

2 Agents' knowledge: a pragmatic analysis

Consider the following knowledge claim: *X* knows *p*. Philosophers have long considered the conditions under which such a claim could be made going back to Plato's *Theatetus*, which analyzed knowledge as *justified true belief*: i.e., *X* knows *p* iff:

1. *X* believes *p*
2. *p* is true
3. *X* is justified in believing *p*

[Gettier, 1963] has shown by a series of counterexamples that this analysis is flawed. Despite considerable effort, no satisfactory analysis of knowledge has emerged that does justice to these or newer counterexamples (largely due to difficulties in defining a satisfactory notion of justification). Knowledge attribution has long been recognized as ripe for a treatment grounded in a more *pragmatic* understanding.

In our analysis of knowledge attribution, *X* knows *p* iff:

1. *X* has *ready access* to *p*
2. *p* is true
3. *X* can *make use of the informational content* of *p* (equivalently, *X* can exercise certain capacities dependent on its knowing *p)*

We retain the truth condition of the original analysis and introduce two new conditions. Condition 1 uses the notion of access to, or easy availability of, the proposition *p*. Condition 3 – which replaces the notion of justification – implies counterfactuals such as "If *X* did not know *p*, then *Y*" – where *Y* is a statement like "*X* is not able to exercise capacity *C*". Since *X* is able to exercise capacity *C*, it knows *p*. (See dispositional [Levi and Morgenbesser, 1964] or functionalist accounts [Armstrong, 1980] for similar notions.)

3 The Case for the Pragmatic Analysis

The following examples illustrate the intuitions underlying our analysis:

1. As I walk down the street, I am asked by a passerby, "Excuse me, do you know the time?" I answer "Yes" as I reach for my cellphone to check and inform him what time it is. This example, due to Andy Clark [2003], shows that we take ourselves to know those items of information that are easily accessible[1] and can be easily used. Clark's example is part of an extended argument for distributed cognition through external tools and memory stores not confined to the inside of our craniums. For our analysis we note that information at hand can be described as information that we 'know' (I could not claim to know my friends' telephone numbers if, on being asked, I were to reply, "I can't remember").

2. A friend wants to buy me a book as a gift. He asks me for my shipping address so that he can send me the book. I direct him to my wish-list at Amazon.com saying, "Amazon *knows* my shipping address". Indeed, the shopping cart software on that site does. When my friend has decided which books he wants to buy, he pays and picks the shipping option. Amazon generates a shipping invoice complete with shipping address. The shopping cart software is able to discharge its functions using that piece of information. I had stored the shipping address on Amazon precisely for such future use. I could store an alternate shipping address and *forget its details*, since Amazon 'knows' it and will ship to it if anyone decides to send me a book at that address. Note that parts of my address could be obtained from Amazon's database (by Amazon.com programmers) by writing specific queries (e.g., "What street does customer *X* live on?"). In a weaker sense, then, Amazon then also knows which of its customers lives in postal code 11205.

3. I have to attend a meeting at the university campus branch located in the city center. With directions for the meeting written down on a piece of paper that I keep in my pocket, I head out the office door. As I do so, my office-mate asks me "Do you know where the meeting will be held?" I answer, "Yes" as I hurry towards the next train. Here, Gricean semantics [Grice, 1975] is at play: *if I said I did not know the meeting's location, I would be misleading my questioner*. This example is crucial in showing that knowledge claims are connected to the pragmatics of speech. To deny a valid knowledge attribution in this case would be to say something misleading or something whose semantic value is considerably less than the knowledge claim.

In the Amazon example the agent is able to use my address to fulfill its functions. An alternative locution would be to say, "Amazon *has* my address", but what purpose would be served other than an avoidance of intentional vocabulary? If it did not ship to the correct address, Amazon could not use as a defense the claim that it did not know (or 'have access to') my address. It was stored in their database and had been used successfully in the past. Amazon is capable of informing a potential customer that it is unable to ship goods since it does not 'know' the recipient's address (or credit card number). If Amazon has access to my address but it has changed in the meantime, then it is natural to say that Amazon *does not know my address since it would not be able to perform the function of shipping books to me*.

When we say, "Amazon.com knows my shipping address", our analysis implies that:

1. Amazon has ready access to my shipping address in its databases.
2. The shipping address is correct.
3. Amazon is able to perform capacities dependent upon its knowing my address (it is able to make use of the informational content of the address).

Furthermore, the relevant counterfactuals are true: if it did not know my address, Amazon's core functionality with respect to its interactions with me would not be achievable; if Amazon did not know my shipping address, it would not be able to send books to me; if Amazon did not know my address, it would not be able to send me a bill; but it is able to do so; hence it knows my address. This kind of analysis is readily extended to other kinds of agents that take actions based on information at their disposal. An artificial agent's actions could be described in much the same way as a human agent's – "The pricebot sent me a quote because it knew my preferences".

Our analysis may be fruitfully contrasted with conventional formal analysis, in which an agent's belief corpus is taken to be the set of propositions that the agent is committed to (the agent answers "Yes" when asked "Do you believe *p*?"). Formally, *p* is derivable using the inference machinery built into that agent's architecture. Thus I could say of an artificial agent that it knows or believes *p* if *p* is derivable from its knowledge base. But to limit knowledge attri-

[1] [Parikh, 1994] has described knowledge as statements that the agent is capable of deriving (from a given set of premises) within some reasonably tractable bounds as opposed to those implied by epistemic closure (if *X* knows *p*, and *X* knows that *p* implies *q* then *X* knows *q*). Parikh's account limits knowledge to those *q*'s that meet the tractability condition (thus providing an analysis of knowledge in terms of the abilities of the agent as we do).

bution to those agents that are capable of deriving the proposition *p* using a formally specifiable inference mechanism would be to put the proverbial cart before the horse. We feel comfortable making the claim that the cat knows a mouse is behind the door though we do not have the foggiest idea of what kind of inferential mechanism is at hand. The cat reveals its knowledge through its actions. Whatever kind of retrieval or inference mechanism is at work, it enables the cat to go about its tasks. Similarly for an artificial agent – it reveals its knowledge of *p* through the ready availability of the proposition in facilitating the agent's functionality. We do not discount a hybrid architecture that employs a deductive database that can infer further information applying rules to a set of stored facts. In that case we would say that the agent in question knows all the facts derivable from its database as well – subject to tractability conditions as in [Parikh, 1994].

Where does an agent's epistemology come into the picture? If an agent elicits information from humans then the responsibility of ensuring the accuracy of the information is the user's. If the user inputs incorrectly, the agent is in possession of false information and we do not make the knowledge attribution. Thus, the software artifact inherits its epistemology from the humans that supply it information and carry out data entry. This should not lead us to think that artificial agents do not have an independent epistemology. Pricebots that read price information on remote web pages acquire knowledge autonomously, by using their file-reading mechanisms, presumably equipped with error checking and validation routines that guarantee it will not read in garbage (the software equivalent of a reliable sensor). The accuracy with which these agents acquire information is a function of their design and the code that runs on them – very similar to human agents, the accuracy of whose beliefs is a function of how well their senses work in conjunction with background knowledge and their reasoning powers.

We would not want to say that an artificial agent knows a proposition if the proposition is simply stored in the agent's knowledge base but is not accessible for use by the agent. In that case, we would say that the information in question is stored in the agent but the agent does not know it, since it is unable to access or use it. Note that when files are deleted from a computer the information ordinarily does not vanish, it simply becomes a target for over-writing. The information is not accessible any more without elaborate recovery methods, and hence the computer's operating system is reasonably enough said not to have access to it any more.

In the case of Amazon, it is possible that not a single human being employed by Amazon knows my address. It is conceivable that when the shipping invoice is printed out by the software, a human clerk will pick it up and attach it to the box of books in question without bothering to check any further whether the address is correct or not. The software has been treated as a reliable source of information with regards to the address – and thus humans might accurately claim that they learned a proposition from a software agent. When a book is purchased, my address has been used by Amazon.com without any human knowing it. What sense would it make to say that Amazon did not know the address? Alternative locutions for describing this functionality of Amazon's would be artificial. What could we say – that Amazon has access to this true information, and can use it? The parallels with knowledge attributions to human agents should be clear. For human agents, on our analysis, are said to know a proposition *p* when we can make such a claim. If I know that 619 times 3 is 1857 but cannot open a safe with this combination, then I do not know the combination to the safe since I cannot open it but I do know the product of 619 and 3.

In making knowledge attributions, there is a parallel between humans and artificial agents. The ease with which we slip into the intentional attribution when it comes to Amazon.com is an indication of this similarity. The intentional stance is used when it is possible to give the best explanations of behavior using it. What kind of behavior would we able to predict? We could predict Amazon's responses to certain queries. For instance, we could say that Amazon knows the ISBN number for *How Green Was My Valley* since it would be able to produce that number on request. But as we have argued, we could also predict Amazon's success in certain tasks – Amazon could demonstrate its knowledge of the ISBN number of *How Green Was My Valley* by shipping me that book and none other.

Condition 2 of our analysis is crucial (as in most analyses of knowledge) since if the shipping address in question were incorrect we would not say that Amazon knows the shipping address. The locution we would employ if Amazon were to use the incorrect shipping address would be "Amazon shipped my books to what it thought (or believed) was my correct address". We would not make the claim that Amazon knows my address if in fact, the address is false (since it is possible to have a false belief).

One way to deny that artificial agents can know propositions would be to ask, "*Who* does the knowing in the case of the artificial agent?" Our response would be that the same could be asked of humans, and in the absence of any philosophically satisfactory analysis of personal identity there is no reason to believe that a stronger condition should be placed on artificial agents. Below, we suggest that the correct legal treatment of artificial agents is to assimilate them to human agents, but without the personhood possessed by human agents. If the same view is adopted on the philosophical perspective, it becomes otiose to ask who does the knowing in the case of an artificial agent – other than the agent itself, of course.

4 The legal doctrine of attributed knowledge

This inclusion of the ready-to-hand in the knowledge of an agent has close and instructive parallels in the legal doctrine of attributed knowledge. Under this doctrine, the law may impute to a principal knowledge – relating to the subject matter of the agency – which the agent acquires while acting on behalf of its principal within the scope of its authority. The scope of the agent's authority refers to those transactions that the principal has authorized the agent to conduct. In some circumstances, knowledge gained by the agent outside the scope of the agency can also be attributed to the agent's principal.

Once knowledge is attributed to the principal, it is deemed to be known by the principal and it is no defense for the principal to claim that he did not know the information in question, for example, because the agent failed in its duty to convey the information to the principal.

The doctrine of attributed knowledge has many applications, and is used generally in civil law contexts in cases where the knowledge of the principal is relevant. For instance, legal consequences attach to principals knowingly receiving trust funds, or having notice of claims of third parties to property received, or knowingly making false statements.

The doctrine has close parallels with our analysis above, which extends the concept of knowledge to include the information that we retain in storage devices – including written documents – that are ready-to-hand. From this perspective, a human agent is akin to a knowledge storage device under the control of a principal. Below, we suggest that artificial agents can be thought of similarly. But first, we explore the basis of the doctrine and its application to the modern company.

4.1 A duty to communicate?

While the doctrine of attributed knowledge is pervasive in the legal systems under discussion, its precise doctrinal basis is still a matter of some dispute [DeMott, 2003].

One explanation of the doctrine relies on the supposed identity of principal and agent, whereby the law sees them as one person for some purposes. However, this theory lacks explanatory power, and does not explain the public policy justification for the rule.

Another explanation put forward for attributed knowledge is that the law presumes that agents will carry out their duties to communicate information to their principals. For example, in the standard practitioner's text *Halsbury's Laws of England*, the scope of an agent's duty to communicate determines the existence and the timing of any attributed knowledge of the agent.[2] Under this approach, the doctrine operates on a pre-existing duty to convey information to deem that the duty has been discharged.

In the US, the common law of agency does not require as a precondition an *existing duty to communicate* the information to the principal [DeMott, 2003]. As Langevoort [2003] points out, the courts' description of attribution as *the presumption that the agent has fulfilled its duty of candor in conveying information* is not correct, since attribution applies *even* where interaction between principal and agent creates enough scope of discretion that no transmission of information is expected (or occurs).

In England the "duty to communicate" has been abandoned as the explanatory basis of attribution of knowledge.[3] Similarly, Australian courts have inferred attribution of knowledge in the absence of a duty to communicate information in cases where the task assigned to the agent included making appropriate disclosures. [4]

We believe the rejection of the duty to communicate is correct on policy grounds To require such a duty in order to attribute knowledge held by agents to their principals would encourage principals to ask agents to shield them from inconvenient information, and would put principals acting through agents in a better position than principals acting directly. Such an approach is also incompatible with modern information management practices within companies, and we discuss why below.

However, the fact that agents are *capable* of communication is important to the attribution of knowledge. In terms of our analysis of knowledge, a lack of capacity to communicate information would render the first and/or third conditions unfulfilled – i.e., that the principal has ready access to the knowledge held by the agent, or that the principal can make use of its informational content. *The capacity to communicate* therefore plays an explanatory role when thinking about how artificial agents fit within this legal schema.

4.2 Attribution of knowledge to companies

A company is a special kind of organization that, in modern legal systems, is recognized as a legal person in its own right. How, then, does a company gain knowledge in the eyes of the law? Apart, possibly, from knowledge gained "directly" by the Board or general meeting of a company, only through the attribution to it of knowledge gained by its agents (i.e., its directors, employees or contractors). By the

[2] Vol. 2(1) (Fourth Edition Reissue) Agency, para 164.

[3] *El Ajou v Dollar Land Holdings plc & Anor* [1994] 2 All ER 685, at 703–4 per Hoffmann L.J. See also [Reynolds, 2001], Article 97(1) at paragraph 8-207.

[4] *Permanent Trustee Australia Limited v FAI General Insurance Company Ltd (in Liq)* [2003] HCA 25 at paragraph 87.

doctrine of attribution, the company is deemed to gain the knowledge that is gained by the natural persons (i.e., humans) engaged by it. [5]

The large modern company illustrates why the "duty to communicate" cannot found the attribution of knowledge. Given the company is an abstract entity, the only way to make sense of such a duty would be in terms of communication to other agents (such as immediate superiors), who are required to communicate it "directly" to the company as embodied by the Board of directors (or general meeting). Since in modern corporations authority to enter and administer contracts is usually delegated to relatively junior staff members, it would be absurd if all the knowledge that had legal consequences for a company had to be communicated upwards in this way. Instead, most information within the modern corporation remains with lower-level officers, and is only passed upwards in summary terms – or when there is some exceptional reason to do so, such as a dispute with outside parties. Abandoning the "duty to communicate" allows the legal system to acknowledge information managed in accordance with modern decentralized practices.

Today the most common way for information to be stored and controlled by low-level officers is by inputting it into the company's information systems. Some of these systems can be queried by senior managers – but it has never, to our knowledge, been suggested that this is essential to the attribution. To what extent could information systems – artificial agents – themselves be treated by the legal system as agents for the purposes of attribution of knowledge?

4.3 Artificial agents as agents for knowledge imputation purposes

In [Chopra and White, 2004], following [Kerr, 1999], it was argued that the legal system should and could extend the legal treatment of human agents to artificial agents, with appropriate modifications. Artificial agents, on this approach, would have a legal status akin to slaves in Roman law – that is, with capacity to enter contracts on behalf of their principals, but without contracting capacity or legal personhood in their own right.

We believe a similar move can and should be made with respect to the imputation of knowledge. On this approach, knowledge gained by artificial agents employed by corporations could be attributed to the corporations themselves, where that knowledge would be attributed to the corporation in the case of a human agent.

Our analysis could be utilized in a legal context for the task of determining what is known by the artificial agent in question.[6]

The scope of the agency would be those transactions that the artificial agent has been deployed to conduct. Not all the agent's knowledge would necessarily be attributed to the principal. For example, an agent could conceivably act for two principals and in accordance with the law of agency, knowledge gained in the course of one agency is not always attributed to the other principal.[7] A natural person could deploy an artificial agent, and in that instance the agent's knowledge would be attributed to the principal in the same circumstances.

Surprisingly, there is a paucity of judicial pronouncements on the possibility of attribution of knowledge held by artificial agents. Some tangential judicial support for such a treatment of artificial agents was given recently, but it was made clear that it did not (yet) represent the law. In the Australian case *Commercial Union v Beard & Ors*[8], the issue arose whether a fact contained in a news clipping, filed in a company paper file, was "known" to an insurer for the purposes of the relevant statute. If it was known to the insurer, the party taking out insurance was relieved of the obligation of making disclosure of the fact to the insurer.

The majority found that a matter could be "known" by the insurer company if it were contained in the "current formal records" of the company. This term appeared to include the minutes of the company's Board meetings. However, the majority held that the extract of the news clipping in question was "not a record of [the insurer] and it was not contained in any file to which officers of [the insurer] were expected to have recourse for the purposes of the subject insurance."[9] As a result, they found that the contents were not "known" to the company for the purposes of the statute.

The minority judge, however, seemed to discount that anything could be "known" to the company merely by being contained in a record, while acknowledging that such a view had its attractions [emphasis added]:

> We were not referred to any authority for the proposition that, in the absence of actual knowledge on the part of relevant officers of a company, the company may, nevertheless, "know" a matter, where the relevant information is contained in a company file. *I*

[5] See *Halsbury's Laws of England*, Companies, 7(1) (2004 Reissue), paragraph 441: How a company may act.

[6] The 'actual knowledge' of human agents is treated by the Courts as self-evident and not needing further analysis: see the five-fold categorization of knowledge in *Baden v. Société Générale*, [1993] 1 W.L.R. 509, 575-76.

[7] On these cases, English and US law take divergent and sometimes confusing approaches: see [DeMott, 2003]; [Reynolds, 2001], at paragraph 8-210; *Permanent Trustee Australia Co Ltd v FAI General Insurance Co Ltd* (2001) 50 NSWLR 679 at 697 per Handley JA.

[8] [1999] NSWCA 422

[9] per Davies AJA at paragraph 63

> *find the proposition an attractive one*. In circumstances, which are undoubtedly common today, where important information relating to the conduct of a company's business is stored in the company's computer system, *from which it may be readily obtained*, the suggestion that such material is part of the company's knowledge is certainly appealing. However…the present state of authority does not permit a finding that the information so stored becomes "known" to the company until it is transferred into the mind of an officer, who is relevantly engaged in the transaction in question[10].

The emphasis on being readily obtainable echoes condition 1 of our analysis of knowledge. Indeed, we suggest that, had the contents of the news clipping been stored in an information system, rather than a paper file, so as to be *readily available* to the human officers conducting the insurance transaction, the result should have been different in the *Commercial Union* case. The prohibitive cost of insisting on cross-checks being made of all paper files before proceeding with any transaction without fear of legal consequences is obviously significantly reduced if those files are held electronically and therefore ready-to-hand to employees of the company generally.

The example suggests that information systems that are mere accumulations of records may not qualify as agents for attribution purposes. We suggest that the knowledge held by artificial agents will only be attributed to a corporation to the extent that the agent permits ready access by other (human or artificial) agents to its contents. In this way, while the duty to communicate is not necessary for the imputation of knowledge, the *ability* to communicate so as to make information readily accessible to others – and not just to passively store information – might well be.

Considering the company as a knowing agent in its own right for a moment, paper files, to which officers are not expected to have recourse in conducting particular transactions, could be equated with the 'dead' information contained within, but not accessible to, an artificial agent – such as information written on a hard disk, but not readily accessible to the user through the operating system without deploying specialized software.

5 Conclusion

We have presented a philosophical and legal analysis of knowledge attribution and suggested that the courts could make use of the analysis when deciding whether to attribute knowledge to artificial agents and principals, such as corporations, employing those artificial agents to conduct transactions on their behalf. In our discussion of Amazon, we did not bother to distinguish between the corporation (a legal entity) and the software agents operated by the corporation. We think that for the reasons mentioned in our analysis above, it can make sense to attribute knowledge to both the artificial agents operated by a corporation and the corporation itself. We also pointed out close and instructive parallels between the philosophical analysis and the legal doctrine. We look forward to the first cases where legally salient information known only to artificial agents is nevertheless attributed to the corporations operating those agents on the basis of the above-outlined principles.

[10] per Foster AJA at paragraph 73

References

[Armstrong, 1980] David Armstrong. *The Nature of Mind*. St. Lucia, Queensland, 1980. University of Queensland Press.

[Chopra and White, 2004] Samir Chopra and Laurence White. Artificial Agents - Personhood in Law and Philosophy. In *Proceedings of the 16th European Conference on Artificial Intelligence, ECAI 2004*, pages 635-639. Amsterdam, The Netherlands, 2004. IOS Press.

[Clark, 2003] Andy Clark. *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford, United Kingdom, 2003. Oxford University Press.

[DeMott, 2003] Deborah A. DeMott. When is a Principal Charged with an Agent's Knowledge? *Duke Journal of Comparative and International Law,* 13:291-320, 2003.

[Gettier, 1963] Edmund Gettier. Is Justified True Belief Knowledge? *Analysis*, 23:121-23, 1963.

[Grice, 1975] H.P Grice. Logic and Conversation. In Peter Cole and Jerry L. Morgan, eds. 1975. *Syntax and Semantics, 3: Speech Acts*, pages 41-58. New York, United States, 1975. Academic Press.

[Kerr, 1999] Ian R. Kerr. Providing for Autonomous Electronic Devices in the Uniform Electronic Commerce Act. Web-only paper at www.ulcc.ca/en/cls/index.cfm?sec=4, 1999. Uniform Law Conference of Canada.

[Langevoort, 2003] Donald C. Langevoort. Agency Law inside the Corporation: Problems of Candor and Knowledge. *University of Cincinnati Law Review*, 71:1187-1231, Summer, 2003.

[Levi and Morgenbesser, 1964] Isaac Levi and Sidney Morgenbesser. Belief and Disposition. *American Philosophical Quarterly*, 1(3):221-232, 1964.

[Lewis, 1996] David Lewis. Elusive Knowledge. *Australasian Journal of Philosophy.* 74:549-567, 1996.

[Parikh, 1994] Rohit Parikh. Logical Omniscience. In *Logic and Computational Complexity, LNCS 960*, pages 22-29, Heidelberg, Germany, 1995. Springer-Verlag.

[Reynolds, 2001] F.M.B. Reynolds. *Bowstead and Reynolds on Agency*. 17th ed. London, United Kingdom, 2001. Sweet & Maxwell.