Web Search engines and distributed assessment systems*

Christophe Heintz
Institut Jean Nicod, Ecole de Hautes Etudes en Sciences Sociales

I analyse the impact of search engines on our cognitive and epistemic practices. For that purpose, I describe the processes of assessment of documents on the Web as relying on distributed cognition. Search engines together with Web users, are distributed assessment systems whose task is to enable efficient allocation of cognitive resources of those who use search engines. Specifying the cognitive function of search engines within these distributed assessment systems allows interpreting anew the changes that have been brought by search engine technologies. I describe search engines as implementing reputation systems and point out the similarities with other reputation systems. I thus call attention to the continuity in the distributed cognitive processes that determine the allocation of cognitive resources for information gathering from others.

Keywords: web, epistemology, trust, reputation, epistemic authority

1. Introduction

Hutchins' (1995) seminal example of distributed cognitive system is the system implemented in the organisation of a navy boat, where numerous individuals and artefacts have their own specific cognitive tasks that contribute to the overall goal of the system: directing the boat. The distributed cognitive systems I analyse in this paper are made of the community of Web users, who link documents, and a search engine. The overall task of the systems is to enable an efficient use of the cognitive resources of Web users. These systems, I argue, are distributed assessment systems. In these cognitive systems, the cognitive task of Web users (as authors of Web pages) is to assess Web documents, and the cognitive task of the search engine is to compile these assessments and produce a 'usable' representation of the result.

Pragmatics & Cognition 14:2 (2006), **387–409**.
ISSN 0929–0907 / E-ISSN 1569–9943

In the first section, I introduce the problems that one wanting to acquire information from others must face. The cognitive function of distributed assessment systems is then described as providing solutions to these problems.

In the second section, I argue that search engines are reputation systems: they attribute to web documents 'reputation labels' in the form of ranking on given queries (SERPs). Search engines provide visibility to web documents, and this visibility directly determines the trusting behaviour of search engine users. An essential aspect of second generation search engines is that their ranking algorithms take as input the linking behaviour of web-users. The consequence is that search engines together with web-users constitute a distributed cognitive system for the attribution of reputation, visibility, and, eventually, credibility.

In the last section, I compare search engines' procedures for the attribution of credibility with the procedures of scientific institutions for the implementation of meritocracies and conclude that they are essentially similar: both rely in the same way on numerous judgements from the individuals so as to compile a representation of the cognitive worth (in degree and range) of sources of information; both constitute the institutional frame of a distributed cognitive system that organise information. In the end, reputation labels from different reputation systems are compiled by end users. As a consequence, reputation systems influence each others ranking. This contributes to reinforce the continuity between pre- and post-search engines era. I thus question anew the extent to which search technologies of the web change the distributed cognition that inform decisions in information search. I conclude with general considerations on the relations between cognition and technology.

2. Information seeking behaviour and cognition — the advent of the Web

2.1 Consequences of the growth of online documents

The number of documents that can be found on the web is huge and there is moreover a phenomenal growth of the number of web sites, which amounted to 217% between 1998 and 2002. Today, Google claims to be indexing 8 billion pages. This phenomenon is much more than just a change in the medium of communication: producing and publishing Web documents is becoming ever easier,[1] with the consequence that the published documents actually come from very different sources of information. In particular, publishing on the Web is not restricted to recognised experts but is also open to the laypeople. Access to online information is, in principle, not restricted to experts either,

and is open to people that may lack the discriminative abilities that would lead them to read or trust only those documents that worth it. These two facts have caused some intellectuals to worry about the growth of the Web and its usage, especially emphasizing the need to filter the huge amount of information that is so easily accessible (e.g., Eco 2003). The normative and political aspects of questions related to the development of the Web seemed to have caused some dramatisation and over-emphasis of the changes that the technologies of the Web have induced or could provoke.[2] So one goal of this paper is to re-balance a possibly distorted view and emphasise the continuity between pre-Web and post-Web epistemic practices.

The techniques of the Web include two major innovations others than electronic publishing: the hyperlink and search engines. The hyperlink (or simply 'link') enables the web users to immediately go, through a simple click, to the linked document. It thus amounts to a reference, which at the same time provides simple access to the refereed document. This technology leads to the practice known as web crawling: a specific practice for the search of information. Also, links among Web documents give to the entire Web a structure whereby documents are characterised not by their content, but by their position in the structure, i.e. by which documents link to them and which documents they link to. The structure of the web is promptly exploited by search engines that use it to get information about the documents.

Search engines are interactive web sites that respond to queries — in the form of sets of key-words — with an ordered list of URL: the Search Engine Result Pages (SERP). Their role has incredibly increased in the last years. It has increased economically, as is made obvious by the current battle among the biggest software companies for promoting their own search engines, and the enormous sums of money involved. But also, and firstly, its role in epistemic practices has become, within a growing range of population, major, and is still becoming more and more important. By epistemic practices, I refer to the routinised behaviours that aim at acquiring or producing new information. There is no doubt that searching for information through the use of search engines constitutes a widespread epistemic practice, as is manifest by the number of queries received by search engines. For instance, at the beginning of 2001, Google answered more than 100 million searches per day. Thus, there is a major change in epistemic practices, and its origin is the implementation of new technologies for managing information.

The success and technical developments of search engines can be understood as a consequence of the growth of information on the Web, including the following facts:

- the quantity of Web documents lies largely beyond what an individual user could ever consult;
- users could not access a Web document without knowing its URL or being automatically directed to it;
- users require some direction in deciding which websites to consult for solving their specific problem.

The reliance on search engines of individuals increased with the number of Web documents, and the number of individuals relying on the Web for the acquisition of information has also increased, hence the importance to analyse and assess the epistemic impact of search engines.

2.2 The stake: Acquiring information from others and cognitive economy

Because there are so many sources of information, the problem arises as to which source one should devote one's scarce cognitive resources. The problem can be framed in economic terms: people searching for information will have to pay time and effort for acquiring the information. Acquiring information from others is often a very good, cost-effective, way to acquire information. Think about asking your way in the street instead of trying all the streets around and making your own map of the area. Attending and understanding others may appear to have a negligible cost with regard to the cost of acquiring the information by oneself. Yet, this cost is much increased when one factors in the risk of being deceived. Imagine you were close to the street you wanted to go to, but some passer-by misled you and sent you in the opposite direction. In this case, trusting somebody else causes you an important waste of time and effort. Also, the cost of asking and attending some answer is rendered much more salient and worthy of interest in 'competitive' situations: Imagine you are in a street with many passers-by. Among them are some who know the area and could inform you well, but some others know nothing about it, and some others wrongly think they know and might mislead you. If you ask to somebody who doesn't know, you will not be misled, but you may miss the opportunity to ask somebody who knows. And you must also be careful not to ask someone that will mislead you. But how can you know in advance which passer-by is a good informant? You may have a few strategies such as 'do not ask someone who looks like a tourist, unless he holds a map in his hand', or 'avoid those who seemed to be in a hurry'. Users of the Web are in a similar situation: they intend to acquire information from others via the online documents, but need to pick and select documents among all the accessible ones. There are numerous strategies for

searching information on the Web and selecting online documents. One such strategy is to start with an already known document, and browse through the hyperlinked documents by clicking on those hyperlinks that seem to answer one's question best. Another strategy is to search outside of the Web — e.g. asking to a friend — for information about where to find specific information in the Web. Often, diverse strategies are used in concert, depending on the goals and beliefs of the information seeker. However, a strategy which is widely used consists in using a search engine. It is a strategy for information seeking that relies on new Information technologies, and which is the focus of this paper.

Strategies for acquiring knowledge from others range from total gullibility to radical scepticism toward what others say. The strategies are to be understood as a way out of the dilemma that people gathering information from others must face. The dilemma is as follows: Before attending others' communications, one wants to be sure that they are worth it, i.e. that one will derive sufficient satisfaction from the time and effort invested in processing the communication (listening or reading, understanding, and drawing the relevant consequences). But how can we know in advance whether a source of information shall provide valuable information? The obvious way to know whether the information is valuable or not, is to assess it. But assessing the information requires paying some cognitive effort that we wanted to spare in the first place by relying on what others say. Likewise, when you wonder whether to buy a yogurt in a shop you may ask first: Is the yogurt worth its price? But you may have no other solution than to pay for it so that you can try it and judge by yourself. So you need to pay the price of the yogurt to know whether it is worth it, and whether you were right to buy it. For information consumption, the problem is even more dramatic, because the assessment of the value of a piece of information does not come with its consumption, as in the case of the yogurt. For instance, attending and understanding someone's communication does not necessarily reveal whether the information communicated is true. The communicator can lie or have false beliefs. So it seems we need to deploy other means so as to check the truth and general worth of what is communicated, and these means can be as costly as finding the information by oneself. The dilemma is therefore: either one consumes the information without knowing its worth, and runs the risk of being completely mislead, or one checks the information communicated, and thus renounces all the advantages of drawing information from others. Attending others' communication seems either hazardous or useless.

As already suggested, the way out of the dilemma consists in checking the likeliness that the source of information shall provide worthy information, rather than checking the information itself. The value of the source

of information guaranties of the value of the information it communicates. This 'method' still includes the risk of spending even more effort checking the source of information than the information itself. For instance, a reductionist standpoint in social epistemology would require knowledge acquired from others to be rationally justified with knowledge acquired by one only. This may be attained with a full and costly check of the reliability of the source of information, but one soon realises that information about sources of information is itself often acquired from others, and the check appears to be a never-ending process. Reductionists thus seem to advocate a non-effective process of knowledge acquisition from others. The unfeasibility of a reductionist check is even more apparent in cases where numerous sources of information are available: should one go through a reductionist check of all the available sources so as to choose the most reliable one?

In actual life, people use strategies that do not comply with the radical criteria of the reductionists. The stake of these strategies is not the attainment of the completely justified certainty that reductionist philosophers would aim at. Rather, the stake is to find information with sufficient guarantees that the information is sufficiently worthwhile for one's purpose, but at a cognitive cost that is minimal.

2.3 Epistemic practices on the Web

Information search on the Web epitomizes the above problems. The Web puts its users in front of a large amount of information out of which they have to discriminate and select worthwhile information. The best strategy for answering the need of web users and maximize the utility gained with the information acquired, is certainly not to browse every Web document so as to eventually select the best ones.

Web users need to rely on information about the information published on the Web, so as to adeptly direct their cognitive effort. However, one sees that the solution can lead to an infinite regress, since one may need also to have information about the information about the information and so on. The information about information must therefore be readily exploitable without further need to go to a meta-level.

What are the actual epistemic practices of Web users — their ways to gather information and acquire knowledge? As mentioned above, Web users heavily rely on search engines. Search engines, indeed, provide information about Web documents that is directly exploitable for deciding on the allocation of cognitive resources.

Using a search engine is a strategy that is cost-effective (note that this does not mean that it is a strategy that is satisfactory in all contexts). In particular using a search engine is a strategy of information search that includes a reliability check of the sources of information. This is due to the fact that search engines' algorithms factor in an assessment of the reliability of the sources (see Section 3 on PageRank). At this point, a key epistemological question arise: is the reliability check that goes with using a search engine satisfactory? In other words, are search engines reliable sources of information about the reliability of the sources of information of the Web? In order to give an answer, one first needs to analyse what information is given by search engine, and how this information is produced, which is the topic of the second section.

2.4 The cognitive function of search engines

I have presented above the difficulties related to the cognitive and epistemic task of acquiring information from others. The problem consists in choosing the source of information to which one will devote further cognitive resources. The best solution to the problem, i.e. the more satisfactory choices for a person, is the one that directs cognitive resources to information that actually has cognitive worth for that person. I have shown that an essential and challenging part of the task consist in finding cost effective ways to acquire worthwhile information from others. Search engine technologies aim at facilitating, or even enabling, a cost effective information acquisition. They arose (first generation, starting in 1990, and second generation, 1997, search engines) at these points in the development of the Web, where the critical mass of information available would make it too costly to search for specific information. In economic terms, this is rendered by saying that search engines answered a demand for effective ways to search information on the Web. In cognitive terms, one can say that search engines fulfil a cognitive function, which was needed for the achievement of a cognitive task of growing importance — searching for worthwhile information on the Web. According to theories of situated cognition (see, e.g., Clark 2003), the cognitive processes that perform information search on the web are distributed between the individual and the search engine: both process representations and transmit information to each other via the key board and the screen. The eventual finding of worthy information is the product of their conjoined processes. Cognitive phenomena that determine the allocation of cognitive resources in people's quest for information on the Web, thus span much further than the head of the individual seeker of information.

The goal of my account of the problem related to the acquisition of information from others was to specify the cognitive function of search engines as elements of information seeking cognition. Because of the limited resources one has, in terms of time and cognitive power to process information, one cannot deal with all the possibly worthwhile information available. I have presented the use of search engines as one of the strategy of information seekers on the Web, which satisfactorily reduce the risk of wasting one's cognitive resources in misleading sources of information. In the next section, I analyse the cognitive part of information seeking processes that is outside of the head of the individual information seeker. Search engines can be simply conceptualised as cognitive artefacts used by used by individuals for information search. But the distributed cognition still extends farther than the human–computer system: it is the whole community of Web authors that is put at work.

3. The Web's assessment systems

3.1 Search engines as meta-information providers

The information given by search engines about Web documents takes the form of lists of URL as answers to queries, which are strings of terms keyed in a text box field. In order to understand the information provided by search engines, one can picture them as enormous registers, where entries include all terms that appear in the Web, sets (i.e. unordered) and lists (ordered) of terms. An ordered list of all the documents of the Web is associated with each entry. In actual cases, some restrictions need to be added to the initial metaphor: for instance, it is the URLs rather than the Web documents that are listed (except for the option 'cache'). Also, the URLs listed are only the URL of indexed Web documents, thus ignoring the mass of documents of the 'hidden Web'. Another restriction is that the algorithms of search engines are such that they will not provide an ordered list of *all* the URL of the indexed sites, but stop the list at, say, the thousandth URL. It is also worth noting that the above information is not stored as such, but the couples (entry, list of URL) are computed 'on the spot', as responses to queries, which are specified calls on entries.

The information contained in these enormous registers is made readily exploitable thanks to the following interactive interface: one keys-in one's query, and the search engines return the list of the first ten (or more) URLs of the associated entry, and the possibility to view the rest of the list of URLs. The hierarchy of URLs makes it easy to consult the first answers and ignore the other

ones. It thus promptly reduces the set of alternatives to consider and makes decisions in information search, processes that are based on manageable data.

With these hierarchies of URLs, search engines distribute ranks in query-indexed SERPs. Ranks provide users with meta-information about the information of Web documents. High ranks are explicitly presented, by Google for instance, as signs of quality and relevance with respect to given queries. Search engines thus provide reputation to Web documents in the same way as academic institutions distribute titles. Academic titles function as guarantees of competence in specific fields, while SERPs' ranks function as guarantees of cognitive worth with regard to specific topics. A SERP rank is tantamount to reputation because it encourages Web users to have a favourable behaviour towards the Web documents having the good ranks. Well ranked documents will be given, on the whole, more cognitive efforts than low ranked one. This means that the ranking functions as indicators of cognitive worth. SERPs give an idea of the worth of a document before the users actually assess the document itself, and ranks are labels of the reputation of Websites. Using search engines is therefore assessing *a priori* the cognitive worth of the Web documents.

There is also some ambiguity in search engines' ranking that is significant inferences and processes underlying assessment of information and their sources. Clicking on a URL sends us to a Web document, and what is actually ranked by search engines are URLs rather than documents. But URLs cannot be equated with documents, since the same URL may host different and evolving documents. The extreme case is when no document at all is stored at a URL, while the search engines had calculated its ranking on the basis of a previously existing document. Everybody using a search engine has at least once been frustrated by such results. What is ranked, therefore, is not the Web document itself, but the source of it as instantiated by the URL. URLs are owned by the people publishing on the Web, so it can be assumed that the source of the documents posted at a given URL, the publisher, is stable across the different documents that are published at this URL. This is what justifies search engines to rank URLs on the basis of the documents that are stored at its place. The inference step is classic in assessment processes: one judges the producer, or the source of information, on the basis of what it has produced or the information it has delivered. And the converse inference is also thought to be warranted: if the source of information has proved to provide good information, then it will continue to do so. Of course, beliefs about sources of information need be constantly updated; and so search engines constantly 'crawl' the web to update their information on documents published and their sources of information. This updating and the relative perennial endurance of web documents even allow search engines to pres-

ent their SERPs as lists of documents rather than lists of sources of information. The link to the 'cache' — which is a copy of the document that was assessed by the search engine and which is stored on the local machine of the search engine — gives a further justification to the inferential step: for instance, if a link is broken, the cache explains why the URL figure in the list in spite of the fact that it provides no information. Remark also that for the rapidly changing documents that are blogs, another procedure is implemented: blog search engines manage information about updating and new posting through 'ping technologies'.

3.2 Distributed cognition on the Web

The hierarchies in SERPs, in order to be informative, must reveal something about the documents ranked. It must be truly useful information about the information contained in the Web documents. The algorithms of search engines do not only peer into each document of the Web in order to abstract information about it. They also use the linking behaviour of Web users. Here is a brief idea of what the algorithms of the Google search engine do. Firstly, it calculates on the occurrences of terms: it is assumed that the more a document contains occurrences of a given term, the more this document is cognitively worthy for the user who specified her interest with the same term in a query. The great innovation of the so-called second generation search engines, however, is to derive information about a document not only from the document itself, but also from the documents that link to it. Thus the famous PageRank algorithm gives a quality mark calculated on the number of links that direct to the document and the quality mark of those documents that make the links. The idea that makes the PageRank algorithm a calculator of quality is that every link made from one document to another can be interpreted as a positive vote from the linking document for its target. In Google's words:

> PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important." Important, high-quality sites receive a higher PageRank, which Google remembers each time it conducts a search.[3]

Note that the interpretation of linking behaviour as positive voting remains controversial. Also, Google itself alternates between 'having value', 'importance'

or 'quality' as best interpretation of 'being linked to'. Let us say, then, that PageRank's output is a judgement of the cognitive worth of documents — as independent of their topic and the interest of the Search engine user. But the main point is that Search engines do not provide such judgements through their own evaluations of Web documents, the judgements are meant to express the ones of the Web users who publish documents with hyperlinks.

With respect to relevance — the fit between the search engine user interests and the topic of Web documents — algorithms take the occurrences of the terms in Web documents as input. But the judgement of Web users is, again, also exploited: the algorithm Hypertext-Matching Analysis takes as input the content of 'neighbouring' documents so as to situate it into thematic clusters of hyperlinked documents.[4]

Because search engines' algorithms take as input the links among Web documents, they rely on the linking behaviours of Web users. These behaviours stem from the users' assessments of the appropriateness of linking to specific documents. Search engines therefore compute query-indexed ranks for Web documents on the basis of all these cognitive processes of individuals assessing Web documents and linking.

Search engines, together with the hyperlinking webpage authors, constitute cognitive distributed systems. Search engines are faced with a cognitive task: producing a representation of the cognitive worth of Web documents. In order to complete this task, they impart the sub-tasks of assessing Web documents to the many Web users that have browsed through the Web, made their own judgment, and created hyperlinks accordingly. The cognition that provides the ranks in SERPs is thus distributed among the search engine algorithm and the many people that create hyperlinks. The distributed cognitive systems — search engines + users — produce and compile multiple assessments: they are distributed assessment systems.

3.3 What the Web's distributed assessment systems do for us

The beneficial functions of the Web's distributed assessment systems are cognitive and epistemic. It sends us back to the first section of the paper. Recall that decisions in information search meet with what appears to be a dilemma: either checking all the information by oneself, measuring the comparative values of pieces of information and eventually select the most worthwhile — this process, when possible at all, requires an enormous amount of cognitive effort that render trusting others worthless and repetitive; or select some information

without checking beforehand its content — but then what guarantee is there that the selected information is sufficiently good?

Web distributed assessment systems provide the setting for implementing a strategy that gets out of the dilemma. The rankings of Web documents are representations of their cognitive worth that are readily exploitable. They can, and they actually do, guide the decisions of information search. Because representations of cognitive worth are closely associated with availability, these representations are readily exploitable and need not call on an third level of information about the information the rankings bear. In other word, what is at work in the exploitation of representations of cognitive worth is not an understanding or deciphering of the content of these representations; what happens, rather, is that availability directly determines users' distribution of cognitive effort without necessarily involving the meaning of the availability. Search engine users, indeed, have usually little knowledge of the processes that underlie ranking. The guaranty of the cognitive worth of a document is not necessarily represented as such in the head of the information consumer, but it is present in the environment as a determinant of behaviour. The cognitive processes of search engine users provide an instance of bounded rationality, which rely on simple heuristics with quickly halting procedures rather than comply with the demoniac rationality of some philosophers and economists (Gigerenzer et al. 1999). Because humans are cognitively limited, they cannot access and comprehend all the information that is available to them. But they seem to rely on processes that manage and economize on their limited cognitive resources (Sperber 2005, Gigerenzer et al. 1999). In particular, humans (as other animals) give cognitive attention to what is at hand, what is promptly present and most salient in their environment. Search engine interfaces exploit this 'lazy' management of information search: Facing SERPs, users will click on the first available links, those that are at the top of the hierarchy and that received the best rankings. Other documents of the hierarchy are given much less attention, if any. The most common strategy consists in browsing the Web documents beginning with the one at the top of the hierarchy and going down the hierarchy, reading the title and excerpts from documents, clicking the most promising ones, and stopping the search as soon as one gets a relatively satisfactory document.

What makes, however, the behaviour of the users of search engines rational? To which extent do these users rationally manage their cognitive resources in their search of information? I hold that Web distributed assessment systems actually enable maximizing the cognitive effect in information search. The reason for this relies on the notion of meritocracy.

First, the act of ranking constitutes the institutional setting where rewards and penalties amount to going up and down the hierarchies, acquiring better or worse ranks. This furnishes a motive for publishing worthy documents. Publishing is an act that (generally) aims at communicating to a large audience. It is rightfully assumed that the URLs at the top of the hierarchies have more chance to get clicked, and the associated Web page visited, than those at the bottom. So the chances of getting to the targeted audience are increased by good ranking in search engines SERPs. If publishers of the Web aim at being read and carefully estimate the role of search engines, then they will want to have high rank. But high rank, as explained, is obtained only through individuals' assessment of cognitive worth. This supposedly insures that only worthy documents will be linked to. Eventually, publishers should aim at publishing worthwhile documents if they want to be read. The motivation to get a high rank in SERPs is not to be under-evaluated. It generated, for instance, a new job specialisation in websites promotion, which includes ranking optimisation. There are many discussions and arguments about how to increase 'artificially' one's web site rank — where 'artificially' means 'without receiving hyperlink references that result from the positive assessment of users of the Web'. It appears also that, at present, commercial web site are much more worried about ranking than other sources of information. Knowledge about the future developments of this situation remains speculative. I would like however to note that the Web remains a huge source of non commercial information on which people increasingly rely. With this increased reliance among different communities, it is likely that ranking will become motivating among these different communities. For instance, while the scientific community is nowadays much preoccupied with impact factor and quotation rate, the intrusion of the Internet and associated search technologies (see, e.g., GoogleScholar and ISI Web of Knowledge) promises to raise the awareness of the importance of rankings. Additionally, the possibility of cheating and other unsatisfactory ranking are a challenge to be met by search engines that have the imperative goal to satisfy users, if only because of fierce competitions among search engines. Attracting users is necessary because search engines attract business clients by showing that they are popular. These facts present the Web as a competitive arena where meritocratic reward systems are being implemented. The merit, I have argued, is measured in terms of cognitive worth, and it is what one participating in the competition should aim at. Of course, this description of the factors which contribute to the implementation of an ideal meritocratic ranking should not blind us to the malfunctions.

The real stake, eventually, is not just the motivation that authors may have to publish worthy document, but the actual implementation of a meritocratic system. With search engines, the meritocratic ideal takes the form of an isomorphy between SERPs returned on queries, and cognitive worth relativised to the topic expressed by the queries. Admittedly, even this ideal is not perfect, since cognitive worth is a value that depends more on the users' interests as defined by his situation and background history rather than on a disembodied topic of interest. It is yet another challenge of search engines software to take this fact into account, since queries can by themselves express but this abstracted aspect of individuals' interests (and ways to personalise searches is being investigated). Yet, there is also a sense in which cognitive worth should include a universal measure: Inasmuch as truth and quality is not dependent on individuals,[5] so cognitive worth has an element that does not depend on people. This universalising idea is also present in the concept of meritocracy: merit does not depend on the person assessing it. Likewise, multiple distributed assessment counts on the fact that what is worthwhile for someone is also worthwhile for the others, and that, consequently for the design of search engines, a hyperlink made by one person is useful information for the others. This assumption seems reasonable — all the more so given that the hyperlinks taken into considerations for queries' answers are the ones of people contemporary to the search engine users and with interests in topics related to their queries. Moreover, the objectivity of the judgement output of search engines is ensured by taking into consideration the judgement of numerous agents — the subjective biases are supposed to cancel each others when added.

The quality of the judgement, on the other hand, is ensured by giving more weight, in the final decision, to sources of information (which are associated with URL) that have already been recognised as worthwhile, and so, trustworthy. Ideally, people would devote cognitive resources to the documents that have the most cognitive worth. This, indeed, would result in a maximisation of cognitive resources, since an investment in cognitive effort would be maximally compensated by the value obtained from processing the document (reading, understanding, memorizing, etc.) Search engines SERPs are intended as a means to achieve such maximisation: meritocratic ordering of cognitive worth leads the users to give cognitive effort to the most worthy sources of information.

Are the new technologies of communication implemented by the Web truly bringing changes in the cognitive processes of information search and selection? In particular, do Web distributed assessment systems implement revolutionary processes? Often, debates on the advantages and drawbacks of the advent of the Web dramatise the changes that it brings with it. For instance,

Umberto Eco (2003) reacts strongly against the epistemic anarchy that, according to him, characterizes the Web. He longs for the use of authoritative filters. Roberto Casati (2003), on the other hand, praises the dynamic of the Web and the voting systems of search engines that initiate feedback on the values of Web documents. However, describing in more abstract terms the principles at work in information selection on the Web shows that there is much continuity and similarity between pre-Web publishing and post-Web publishing.

4. General properties of distributed assessment systems and the specificity of search engines

4.1 Other reputation systems

Here is the characterisation of Web distributed assessment system we have arrived at: The systems produce representations, which are assessments of the cognitive worth of Web documents. The production involves not only the algorithmic computations of the search engines, but also the mental computations of those who surf the web and produce Web documents with hyperlinks. We hold that scientific institutions have developed similar complex processes that seek to optimize both quality and relevance of the sources of information to which scientists dedicate time and effort.

The individual scientist cannot do research without knowing what is going on in her field. This implies knowing which are the controversial articles, which are the ones he should take for granted, which are the ones he needs to read because of their relevance for his work and which are the ones he can skip because of their irrelevance or poor quality. A scientist needs to rely on representations about these cognitive productions; he needs to assess their cognitive worth, and in order to do so, he uses reputation. In practice, researchers — as Web users — will devote their cognitive resources to the sources of information that are readily available. These may be the scientific journals in their library, the articles that appear to be pre-eminent because often quoted, the authoritative scientists that are invited as 'guest speakers' at numerous conferences, etc. One can thus see that the scientific practice of search for information from others is not much different from the practices on the Web.

What is the rationale for devoting cognitive resources to readily available information? The cognitive foundation of this behaviour is straightforward: scientists, as lay people, work under sever cognitive constraints and limits, they just cannot read and understand the entire scientific production, but need

to limit themselves to what is worthwhile for them. How do they determine what is worthwhile information for them? We are again facing the problem of information search decision making, and the solution remains the same: cognitive attention is given to salient and available information. The question that follows is already known: What makes this behaviour rational, adapted to the exigencies of scientific thinking? And the answer is as above: because the institutional setting makes it rational. And this institutional setting, again, is formed by a distributed assessment system.

The production of evaluative representations in science involves a distribution of cognitive tasks among numerous cognitive agents and processes. For instance, the evaluation of a job applicant's credentials is a representation that is determined by several factors including teachers' past evaluations of the applicant, his past employers and journal's referees (since publications constitute credentials). From one's education to the peak of one's career, expertise is constantly assessed. Grant applications, submissions of papers and job applications are part of scientists' everyday life. The evaluation of a scientist could not be brought about by a single person: every scientist go through a very large number of assessments, which stretch over a whole career, and may require different kinds of very specific expertises. Thus, the evaluative process in science is fundamentally distributed. The distributed aspect of scientific evaluation is even more apparent in the assessment of research results. The authority of an article crucially depends on its reception by the scientific community. It is through the judgements of other scientists that representations of the quality, relevance and importance of articles emerge.

The processes through which scientists and institutions assess research relies on a positive reputation system, i.e. a set of procedures that attach to actors or objects an attribute "that signals that they are more likely to be desirable for some sort of interaction than those without the attribute" (Whitmeyer 2000). In scientific practice, the positive attributes include holding a Ph.D. and research positions in more or less prestigious institutions (for people), being published in well known scientific journals and being widely quoted (for articles), or being ranked in the top first universities or most active research lab (for institutions). The pervasiveness of the reliance on indicators of quality is made apparent in Boix Mansilla and Gardner's sociological research on the assessment of scientific research:

> Faced with the task of making their assessment criteria explicit, researchers typically referred to indirect or field-based measures of quality. They pointed to indicators such as the number of accepted patents, publications, devices,

> and citations stemming from the work; the prestige of the universities, funding agencies, and journals in which it is placed; and the approval of peers and a broader community. "Simply counting things are easy answers as far as I'm concerned," claimed Jonathan Rosen, Director of the Office of Technology Implementation at CIMIT. "How many patents have you filed? How many patents have been licensed? How many new companies have been started? How many Science papers? How many Nature papers?" Field-based measures of this kind […] [rely] on social procedures of peer review, inter-subjective agreement, and ultimately consensus as generators of acceptable insight. Our subjects […] described these criteria as the standard way — however flawed — in which the quality of interdisciplinary work is determined at the forefront of knowledge production today (Boix Mansilla and Gardner 2003).

The positive reputation systems of science allows scientists to know whom and what to defer to, whom to trust *without further investigation*: it allows single elements to *reduce the cost* of trustworthiness assessment orienting their decisions. The positive reputation system of science tags sources of information as worthwhile: these sources of information are more likely to be desirable for cognitive interactions leading to fruitful scientific research.

The similarities between Web and scientific distributed assessment system are not so surprising. In fact, the reputation system implemented by search engines is inspired by citation counts of scientific articles — an assessment procedure developed by Garfields in the 50's. Links in the Web and citations in the scientific literature are interpreted similarly as expressions of cognitive worth. Yet, the existence of distributed assessment systems is not restricted to these two cases. Distributed assessment systems are more widespread and comprehensive: in scientific practice they also include the continuous assessment of a life span careers in addition to cite count of publications. When one goes to a medical doctor, one still relies on some distributed assessment system that issued the label of 'medical doctor'. Information about material goods is likewise produced by distributed assessment systems, at least, insofar as consumers contribute to the reputations of brands. And the reputations of brands, in turn, determine consumers' choices. On the Web, commercial Web site such as e-bay and Amazon also implement, with great success, distributed reputation systems.

4.2 General properties of distributed assessment systems

The existence of distributed assessment systems for information search can be explained functionally: they answer to an essential cognitive need — choosing

to which source of information to devote cognitive effort. Distributed assessment systems perform a cognitive task — producing adequate representations of worthiness — that could not fruitfully be pursued by isolated individuals. It would be interesting to investigate further into the generation of the institutional frameworks that produce distributed assessment system. Let me just note, however, that the core institutions of such distributed assessment systems are reputation systems: they process information from multiple sources and produce a representation of cognitive worth. Such institutions can be found in Academic institutions and other reward systems. One may even see 'institutions' such as 'gossiping at the end of Sundays' religious rituals' as implementing reputation systems: gathering and redistributing (after some processing!) information about the reliability and trustworthiness of others. Of course, the functional aspect of these institutions is not by itself the cause of their existence, but it should certainly enter the economic and social histories of their emergence.

So what is new with the Web's distributed assessment systems? The function these systems implement is not new and the principles of the processes implemented neither. The change is to be found in the way Web reputation systems operate: they are *automatised*. It is difficult to pin down what exactly this automatisation brings about. It certainly brings a cognitive power that was never attained before. Few institutions, indeed, are able to centralise and process as much information as search engines' algorithms. This cognitive power is also manifest in the responsiveness of the system. It is continually and rapidly updated. A rapidity that proves essential for grasping the dynamics of the Web. So a change in cognition induced by technologies of Web search is certainly a quantitative jump in cognitive power. Qualitative change may come out of the quantitative growth. At the moment, however, I find it difficult to have a dichotomy of pre-Web and Web eras in terms of with vs. without filters, or dynamic vs. expert based assessment — as in the Eco — Casati debate.

The characterisation of the role and place of search engine technologies in the Web's distributed assessment systems allows pinning down some of its essential characteristics as common to any distributed assessment system for information search. These are:

– The distributed and cognitive aspects of these systems: the systems deal with information, and the processes are distributed among multiple agents, who assess and issue multiple value judgements. The systems qualify as distributed cognitive systems. As assessment systems, they rely also on inferential assimilation of representations on the value of pieces of information and the worthiness of their sources.

- In order to solve the problem linked to the cognitive management of cognitive resources attributed to others as sources of information, the output of the system needs to combine saliency and availability within their representational format. The representational medium use a 'less is more' effect in cognition (Goldstein and Gigerenzer 1999), which means that individual cognition benefit from having less information (a short list of URL more reliably direct cognitive attention than a large set of URLs coupled with detailed accounts of their worthiness). The distributed assessment system structures the environment in such a way that there is a *fit* between the environment and the cognitive processes, which automatically direct attention and effort towards salient phenomena (Sperber 2005). The fit, in our case, is between naturally processed because available, and cognitively worthy — it leads to the maximisation of cognitive resources in one's search of information from others.
- A central problem of distributed assessment systems is to maintain this fit: make sure that the most available sources of information are actually the most worthwhile. This is the condition for the maximisation that is associated with meritocracy: the ones who are listened to are the ones that deserve the most to be listened to. Thus, the worries about distortion and unintended effects should be central issues for reputation systems. One can promptly note the incentive to 'cheat' reputation systems, i.e. bypass the intentions and goals of the system so as to make use of its actual functioning. But there is an opposite interest for the systems themselves that need to keep the cheats away so as to maintain their reliability and efficiency (and thus their existence). Maybe an unintended effect of search engines is to further the preferential attachment behaviour that structures the Web. Preferential attachment is the fact that links are mostly directed towards those URLs that have already been linked to. Since search engines direct attention towards those latter sites, they increase the probability of further linking to them. The unintended consequence is an over hegemonic structure and a lowering of the chances of newcomers to get 'chosen' and linked to.

A last property of distributed assessment systems is relevant to the question of change brought by technologies in the distributed cognition of communities. The representations' output of these systems, reputation labels, can be, and actually are, processed in combination by the end users. This means that I can use the results of a search engine for directing my attention and efforts, but I will also generally double check the results with some further information on the sources of information. For instance, I may use my knowledge of the prestige of

authors when choosing among the ten first results of a SERP. Furthermore, this prestige is the representational output of some other distributed assessment system, such as the after-a-conference-day-gossiping or the prestige of their institutional affiliation (which may itself be the results of some (distributed) evaluation of public educational institutions). Because all these representations of worth are finally combined in the head of end users, they eventually influence each other. The hierarchy established by one reputation system feed in the others' systems' hierarchies. The assessments that individuals produce ultimately operate across reputation systems: when the assessment is taken into account by a system, then it shall (partially) determine the behaviour of some users that will produce further assessments taken into account by other systems. Distributed assessment systems operate across systems.

With the advent of new publishing technologies and media, with the creation of technologically supported reputation systems, the cross-operability of distributed assessment systems is fully effective. At first, it is the old hierarchies supported by already existing reputation system that inform the yet-to-be-built hierarchies. This supports the thesis that there is much continuity in the cognitive decisions managing the allocation of cognitive resources. And indeed, a simple look at the current history of prestigious sources of information shows that no real revolution has taken place.

I have emphasized the continuity between pre- and post- search engines for three different reasons:

Firstly, I wanted to balance discourses which assume that the Web introduces dramatic changes. Secondly, analytical abstraction reveals the hidden structure common to all distributed assessment systems and thus provides a template through which actual changes come out. The existence of important local changes cannot be denied. I have noticed, for instance, that marketing strategies have to face very new challenges introduced by the new stake of 'having one's web site visited'. With search engines, reputation may also free itself, to a certain extent, of the biases introduced by the interests of those institutions that implement their own reputation systems that favour certain criteria rather than others. This is because the Web combines information coming from different kinds of sources, that would not have communicated before. Also, search engines implement a radically new way of carving information: while information is traditionally structured in hierarchical typologies (e.g. discipline, subdiscipline, and field), search engines structure it with key-words and sets of key-words.

The third reason why I have emphasized continuity has to do with the relation between technological innovation and cognition. Often (as in a Marxist

perspective) technological innovation is shown to be the source of changes, social or others. But the converse determination, the social and cognitive determination of technical innovation, should not be ignored. Analyses of the complex network of historical causes and determinations between technology and cognition have focused on different causal chains, as for instance:

Technology → practices (using cognitive artefacts) → (situated) cognition
Technology → modes of production → social structure → (distributed) cognition

But any technological innovation takes place within society at a given time and place. Technological innovation is thus itself situated in a given culture and social structure. Before it can act upon culture and society, an innovation arises from the current beliefs and goals of a community. In order to be successful technological innovations must 'fit in' the social and cultural situation. What 'fit in' really means is maybe the most fundamental question of the history, sociology and philosophy of technology. In this particular essay, I have introduced the cognitive dimension of the 'fitting in', viz. the need to manage cognitive effort in information search. Moreover, I controversially suggested that the existence of distributed assessment systems is pervasive across cultures — i.e., distributed assessment systems would be a universal cultural answer to a universal cognitive problem. Of course, this does not mean that there is a unique cultural solution to the problem of information search. Distributed assessment system is just an abstract notion, which is implemented in many different ways. Web users + search engines would be just another avatar of this cultural universal. But the notion of distributed assessment systems sends back to some basic facts of social life: there are incompetent people and there are liars. But rather than simply avoid collaborating and trusting others, humans have found another solution, which paradoxically implies even more trust and collaboration: humans interact and collaborate in order to solve collectively the problem of information search. Here, the focus is on the causal relations that go from the human nature, especially in what concern human's cognitive apparatus, to culture, which include technical innovations.[6]

Notes

* I thank Roberto Casati, Gloria Origgi, Dario Taraborelli and Dan Sperber for their help and encouragement in the general project of developing an analysis of the epistemic impact of the Web. Gloria Origgi and Dan Sperber also commented on an early version of this paper.

1. Technological advances make it easy to create documents: there are user-friendly editors of html, and the knowledge of the basic skills for web-publishing is spreading; acquiring a domain on the Web is much facilitated and the prices very affordable, when it is not free. Of growing importance are the means of publishing without going through the html and transfer protocols: blogs and the Wikis are such means, which are more and more widely used (for instance, it is technically very easy to write one's article on Wikipedia).

2. Questions such as: shall we promote the development of the Web? How and for which purposes? What means should be used to manage Information on the Web? Shall we implement censorship? — have led to a divide between Internet enthusiasts and Internet skeptics.

3. From Google's web site: http://www.google.com/technology/, accessed on 1 July 2005. For an account of the PageRank algorithm and its underlying mathematics, see Brin et Page (1998).

4. It is difficult to know what the algorithms of search engines exactly do, since they are kept secret for commercial reasons. Supposedly, the Hypertext-Matching Analysis factors in the text that link to the document in question. It may also use some cluster analysis: the analysis of closely hyperlinked sets of documents.

5. Truth and aesthetic claims contain a claim of universal validity.

6. The causal relation that goes from the nature of mind to culture has been emphasized by, e.g., Sperber (1996) and Tooby and Cosmides (1992).

References

Boix Mansilla, V. and Gardner, H. 2003. "Assessing Interdisciplinary Work at the Frontier. An empirical exploration of 'symptoms of quality'". In *Rethinking Interdisciplinarity*, http://www.interdisciplines.org/interdisciplinarity.

Brin, S. and Page, L. 1998. "The anatomy of a search engine. http://www-db.stanford.edu/~backrub/google.html.

Casati, R. 2003. « Ce que l'Internet nous a appris sur la vraie nature du livre ». In G. Origgi and N. Arikha (eds), *Text-e : le texte à l'heure de l'Internet*. Paris : Editions de la Bpi Paris, 51–75. [http://www.text-e.org]

Clark, A. 2003. *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford : Oxford University Press.

Eco, E., 2003. "Auteurs et autorité" In G. Origgi and N. Arikha (eds) *Text-e : le texte à l'heure de l'Internet*. Paris : Editions de la Bpi Paris, 215–228 .

Goldstein, D.G. and Gigerenzer, G. 1999. "The recognition heuristics: How ignorance makes us smart". In G. Gigerenzer, P. Todd, and the ABC Research Group, *Simple Heuristics that Make Us Smart*. New York: Oxford University Press, 37–59.

Hutchins, E. 1995. *Cognition in the Wild*. Cambridge, MA: The MIT Press.

Sperber, D. 1996. *Explaining Culture: A Naturalistic Approach*. Oxford: Blackwell.

Sperber, D. In press. "Modularity and relevance: How can a massively modular mind be flexible and context sensitive". In P. Carruthers, S. Laurence, and S. Stich (eds), *The Innate Mind: Structure and Content*. New York: Oxford University Press.

Sperber, D. and D. Wilson 1995. *Relevance: Communication and Cognition*, 2nd ed. Oxford: Blackwell.

Tooby, J. and L. Cosmides, 1992. "The psychological foundations of culture". In J. Barkow, L. Cosmides, and J. Tooby (eds), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* New York: Oxford University Press, 19–137..

Whitmeyer, J.M. 2000. "Effects of positive reputation systems". *Social Science Research* 29: 188–207.

Author's address

Christophe Heintz
1 bis av. Lowendal
F-75 007 Paris
France

christophe.heintz@gmail.fr
http://christophe.heintz.free.fr

About the author

Christophe Heintz is working at the Institut Jean Nicod on the relations between cognitive and social factors of knowledge production. His research is focused on methodological questions in science studies, on the study of mathematical and anthropological knowledge, and on the role of the Internet in knowledge production. He has been trained in mathematics and philosophy at the Universities of Paris and Cambridge, and currently completing his Ph.D. at the Ecole de Hautes Etudes en Sciences Sociales (Paris).