INNOVATION JOURNALISM
VOL.6 NO. 7 * NOV 11 2009 * ISSN 1549-9049
www.innovationjournalism.org

Digital Identities and Journalism Content

How Artificial Intelligence and Journalism May Co-Develop and Why Society Should Care

Noam Lemelshtrich Latar

Sammy Ofer School of Communications
IDC Herzliya
Israel

David Nordfors

VINNOVA-Stanford Research
Center of Innovation Journalism
Stanford University

The Innovation Journalism Publication Series
VINNOVA-Stanford Research Center of Innovation Journalism
Wallenberg Hall, Stanford University

Digital Identities and Journalism Content

How Artificial Intelligence and Journalism May Co-Develop and Why Society Should Care

Artificial Intelligence (AI) algorithms are changing professional journalism and related academic research dramatically. AI is penetrating journalism's pillars: content (through automatic content analysis in all formats), and advertising (by scientific measurement of real consumer attention and targeting ads per user personality). Both content and advertising will change significantly.

The interactive nature of the new media will permit, for the first time, accurate measurement of the real attention consumers of media give to journalistic content, employing scientific methods. Advertisers will demand full validation of consumer ratings. Existing measuring methods will vanish. Advertisers ROI (Return On Investment) will determine the fate of advertising funded journalism companies across all media formats.

New ways to measure consumer attention and behavior, such as 'engagement' and 'behavioral targeting,' are becoming the new buzzwords describing deeper consumer involvement with content across multiple personal dimensions. New AI algorithms are being created that will allow automatically deciphering and tagging content to enable search engines to seek new, practical knowledge. Video, audio, images and texts are being converted to mathematical formulations that lend themselves to automatic 'knowledge discovery analysis' without human intervention.

AI engines will be used by media companies to search customers for content interests, automatically. Dependence on gaining measurable consumer attention can be expected to induce journalists in all media platforms to adjust content to maximize consumer attention and advertising dollars. New business models will be needed to reduce the intrinsic risk to journalistic freedom that the new methods will induce.

In this paper we shall describe the global efforts in devising universal standards for the management of digital identities and how artificial intelligence will be used to automatically annotate journalistic content. We shall describe the new concepts being used to increase consumer real attention to media content and describe the architecture of an AI engine that will target content according to consumer personalities. The consequences of these developments will be discussed.

> “*It is very important that when we click*
> *we click without a thought that a third party*
> *knows what we are clicking on…*
>
> *I came here to defend the internet as a medium…*
> *To allow someone to snoop on your internet traffic*
> *is to allow them to put a TV camera in your room,*
> *except it will tell them a whole lot more about you*
> *than a TV camera.*
> ”
>
> *Tim Berners Lee, Inventor of the World Wide Web*
> *UK House of Parliament, March 2009*

1 Introduction

Berners-Lee's attempt to defend the Internet against ISPs that profile customers by ‘snooping’ echoes similar attempts by the scientific pioneers of nuclear energy, following Hiroshima, to control and limit the use of nuclear energy to the peaceful service of humanity.

The issue involves trade-offs between privacy and security, between privacy and convenience, and between privacy and business opportunities. The actors include isolated hackers and criminal networks, respectable companies pursuing market-driven business trends, and NGOs.

Survival of the world economy in the age of information, and the avalanche of digital data, depends significantly on introducing automatic decision-making procedures to maintain competitiveness, while preventing organizational collapse due to information overload. Survival of global conglomerates such as Google, Yahoo and the ISPs (but also any other organizations handling large amounts of data), will depend critically on their ability to improve their search engines by adding smart algorithms employing AI tools that can analyze hundred and thousands of variables simultaneously in an uncertain world.

Organizations that cannot provide a meaningful answer within a click or two may not be competitive on-line and may even cease to exist. This will require strong correlation between the consumer ‘total being,’ as expressed by a digital format, and the products or information sought. This ‘total being’ may consist of the person’s cognitive attributes, temporal mood and the contextual environment at the time of the 'click', even the individual’s genetic code,

A major debate is occurring on the issue of privacy, amid calls for legal enforcement of ‘transparency and control’ by consumers. Expecting consumers to understand the complex, long- term issues involved before clicking the "I Agree" button on so called 'contracts' (frequently appearing in web sites as a "pre-condition" to press the "continue" button), is evading the issue. Several conceptual frameworks are being used to describe the vast amount of activity already taking place to adapt products to consumers' personalities: "Engagement," "Behavioral Targeting," "Contextual Targeting," "Personalization" and others.

In this paper we focus on 'journalistic content behavior targeting' vs. Digital Identities (DI). Targeting journalistic content via digital identities has special significance as information empowers people. Journalistic content comprises multimedia-format news of information and cultural products that affect consumers' personal development and their access and use of societal resources. Letting digital identities filter content automatically, to affect content and determine access, will engender serious social consequences far beyond privacy considerations.

Journalistic content is undergoing major changes due to the new, technological interactive platforms being introduced that make media content available continually, everywhere. Until recently, the mass media for distributing content were controlled by the same companies that produced the content. The traditional business model for news and entertainment includes controlling both the medium and the content, and bundling them.

But with the Internet, a new generation of incumbents is arising in the media industry. Companies such as Twitter, Facebook or Google consciously avoid producing content. Journalism is separating from the media[1]. The latest generation of producers of journalism is no longer involved in the processes or infrastructures of mass communication. They focus on producing content and publishing on-line, making it available through the infrastructures of the new content-neutral media entities. The Huffington Post and TechCrunch started as blogs. They are today large and important publications, without controlling infrastructure for spreading content other than their Internet domains.

This begets new major avenues of competition between new and traditional media companies clamoring for consumer attention.

Both old and new media companies apply 'attention business' models, generating and brokering the attention of their audiences/users, typically by selling ads. People working for media organizations earning revenues on ads may therefore be seen as 'attention workers'[2]

[1] Nordfors, David 2008, "Separating Journalism and the Media" EJC Magazine 4 Dec 2008, European Journalism Centre
http://www.ejc.net/magazine/article/separating_journalism_and_the_media/

[2] Nordfors, David 2006. "PR and the Innovation Communication System", Innovation Journalism Vol.3 No.5. (2006), http://www.innovationjournalism.org/archive/INJO-3-5.pdf , also published by Strategic Innovators (July - Sept 2007, Volume I | Issue 3)

In traditional media, advertisers pay for audience attention by paying for co-exposing themselves with content produced by the media company on the medium the media company controls. This could be an entertainment or news item in a newspaper or on TV, or within a space on the Internet under the control of the media company. The market decides the price—an ad spot in a popular TV show is more expensive than in a less popular one.

In the new media, advertisers pay for user attention in other ways. Google charges advertisers for co-exposing themselves with the search results for keyword searches. Again, the market decides co-exposure cost. While writing this essay, the word "Ford" had a maximum cost per click (MCPC) of $5.44, "Ford Mustang" had a MCPC of $13.11,while "Mesothelioma" had a MCPC of $375. Mesothelioma is a rare, incurable form of aggressive cancer caused predominantly by asbestos exposure and therefore greatly interests, for example, lawyers who can help people carrying the disease bring lawsuits against the parties responsible for asbestos exposure.

In return for attention, consumers get proprietary editorial content from the traditional media companies. From the new media companies they can access content published by others, and often access to free tools from the new media company. Google offers search engine and other tools free, like bloggers, word processing and email. Facebook offers free tools for social networking that produces content.

This new battleground and new 'content tools' force new and old media organizations to innovate constantly and create new media experiences to achieve maximum consumer engagement. Virtual worlds, computer games and alternative-reality games (ARG) are being introduced into media content to enhance user experience and thus gain attention.

Publishing journalistic content, with a business model based on generating and brokering the audience attention through the medium, therefore becomes increasingly complex. Journalism no longer competes only with journalism and entertainment in other traditional mass media. It now competes with Internet searches, amateur content (family photos, social interaction between Internet users)—in effect, anything that catches people's attention on the Internet, with much more to come.

The Internet competes even better in media and content supported by attention business models, e.g. journalism paid by targeted ads. In traditional media, advertisers pay for the probability of catching consumers' attention, corresponding to another probability: consumers acting on the ad.

Traditional media companies spend hugely to measure readerships, estimating their sizes and the probabilities of catching their attention and creating statistics and probabilities that will appeal to advertisers. On the Internet, on the other hand, the new media companies can offer their customers—content producers and advertisers alike—not probabilities, but hard data on which user looked at what, where, when and for how long. Advertisers will know if a reader clicked on their ad or not. So while traditional media sell attention probabilities, the new media companies sell actual attention. Ads in traditional media are indiscriminate, broadcast to all consumers and costs the same, regardless how many people pay attention or act. Ads in the new media can be targeted in many ways, and paid for per user who clicked on them.

The Internet enables ‘contextual advertising,’ where advertisements shown to each user in the audience are selected and served by automated systems based on content displayed to the user.

The vast possibilities of monitoring users and adapting content and ads to individual users are revolutionizing content, media and marketing. In a digital-interactive world, marketing must account for every dollar spent on buying media. The ROI on every dollar invested in advertising and targeting content to the right consumers is becoming a science, a formidable driver for the accelerating development of advertising, media and content. In 2007, global spending on advertising was estimated at $385 billion [3], equivalent to the entire 2008 GDP of the world’s 26th largest economy according to the World Bank[4]. According to a 2007 report by Piper Jaffray, the 2008 global market for contextual advertising was projected to $35 billion, roughly a tenth of the total world market for advertising, and expanding rapidly.

This is expected to affect journalistic content significantly in all modalities and may revolutionize the journalism profession and its academic research.

Journalism must adapt to new advertising models and investigate business models other than ads that provide the costly practice of journalism with competitive edge to its on-line rivals for generating and brokering user attention. One example of alternative business models: on-line conferences, generating revenues for news organizations in Silicon Valley that cover the innovation economy, such as VentureBeat or TechCrunch.

Contextual advertising is only starting. In simpler forms, it merely matches ads and web pages that share particular keywords. This can give unwanted results for advertisers. For example, the Seattle Times published in 2004 the story “Chance of Mount St. Helens [volcano] Eruption Grows.” The on-line story was accompanied by the ad “Cheap Hotel Rooms in St. Helens.” The algorithms for contextual advertising did not understand that the right time to sell hotel rooms is not when the risk of a volcanic eruption is in the news story.

There is obviously much to gain by developing contextual advertising. Optimally, it should combine the content on a web page with an ad, to match the present state of mind of the individual user studying the page, such that the user is inspired to engage in a transaction with the advertiser. Profiling users is therefore a high

[3] Wikipedia Aug 29 2009. http://en.wikipedia.org/wiki/Advertising#cite_note-2 "Global Entertainment and Media Outlook: 2006–2010, a report issued by global accounting firm PricewaterhouseCoopers". Pwc.com. Retrieved 2009-04-20.

[4] "The World Bank: World Development Indicators database, 1 July 2009. Gross domestic product (2008) http://siteresources.worldbank.org/DATASTATISTICS/Resources/GDP.pdf

priority, potentially optimizing desired and voluntary interaction between people that will increase wealth and well-being for everybody involved. But it brings risks of privacy infringement and can enable various individuals and stakeholder groups to gain benefits by harming others. Control over and access to such data are therefore rapidly growing, important existential issues.

Targeting new content per consumer digital identity will require AI engines to analyze this multi-dimensional content along attributes of the engaging experience and the 'total being' of the consumer—relate the human DNA, the content DNA and the context DNA (attempts to identify successful music DNA and literature DNA already exist and will be described).

Research in biology, genetics and psychology that explore and identify links between individuals' genetic codes, their cognitive attributes and pro- and anti-social behavior is merging with data mining relating to Web 2.0 social-network activities aimed at consumer profiling. Digital Identities will integrate a person's genetic code with data derived from web clicks. People will pay with privacy for social networks benefits. Digital dossiers never go away.

In this paper we describe the process of applying innovative Artificial Intelligence algorithms to analyze journalistic content in all of its formats—text, video, audio and still images—to annotate(tag) content automatically.

We describe the global efforts to create unified digital-identity standards to every individual on the planet and use AI engines to target annotated content automatically vs. individuals personality traits, possible including genetics. This 'journalistic content targeting' may alter the traditional foundations of the journalism profession, not necessarily for the better, and will require new business models for the survival of this important profession.

In the first section of the paper we describe the formation of digital identities. A brief outline of binary coding will be provided. New global standards for digital identities and the use of social networks, genetics and virtual worlds for creating DI will be discussed.

In the second section we describe new, innovative AI research being used to code and annotate (tag) video, images, text and audio content automatically. Scientists are converting journalistic content to mathematical formulations ('signatures') to understand content and context.

In the third section we describe the recently popular concept of media engagement and its derivatives—behavioral targeting (BT), contextual targeting, and how AI is being used in social networks to target content and ads.

In the forth section we describe an AI engine that is being employed by governments for automatically allocate resources and access to services to their citizens based on some initial personal data and how a similar 'engine' can be used to filter and target journalistic content based on the digital identity of the consumer in order to maximize the ROI of every dollar spent on advertising.

2 FORMING DIGITAL IDENTITIES

2.1 Binary Code, Digital Identity and Data Mining

The digital information age was created following the ability to translate essentially all information, including all recordable human actions, into binary (dual) code[5]. Data translation into binary code has existed for hundreds of years. As early as the 16th century, the mathematician-philosopher Francis Bacon (1561-1626) invented a binary code to cipher the language comprising two letters—A and B—as described in his book *Advancement Of Learning*. Binary logic fundamentals laid down by the German mathematician-philosopher Gottfried Wilhelm von Leibniz (1646 – 1716) were cornerstones for developing logic theories by British mathematician-philosopher George Boole (1815 – 1864), among the founders of modern logic.

Boolean logic solves complex problems by dividing them into units based on simple structures; the solution follows a sequence of binary questions and choosing between two possibilities at every stage of the solution.George Boole's logic structures were translated into electric circuits that led to the development of the modern computer, the basis for the information revolution.

Samuel F.B. Morse (1791 – 1872), the American portrait painter from Massachusetts, invented the telegraph and the method to send messages based on binary code, dot and line. Morse improved the code whilst taking the statistical structure of the language into consideration. Morse and Bacon's codes were later replaced by the binary 'bit,' which represents an electrical pulse in a sequence of electrical pulses that activate the modern computer.

Communicating, storing and processing data has gone from craft to science thanks to Claude E. Shannon, who in 1948 published his classic paper "A Mathematical Theory of Communication," which gave birth to the field of Information Theory. Another notable contributor is the Nyqvist-Shannon sampling theorem, which says that any analog signal can be sampled and digitized. The tremendous importance of the sampling theorem stands clear today, when music and pictures—recently considered obviously analog—are today regularly sampled to the extend that we think of them as digital objects.

Shannon also proved that a digital signal can always be transferred through a noisy channel without information loss. This is as important as the sampling theorem and enables digital communication. It shows us the importance of coding the digital data suitably, whether compression, error correction, or suchlike.

With this, we entered the digital age, where information is sampled into the digital domain, to be stored, processed and communicated with increasing ease and efficiency.

[5] Binary code is the system of representing text or computer processor instructions by the use of the Binary number system's two-binary digits "0" and "1".

Nicholas Negroponte of MIT calls the single digit—the bit—"the smallest atomic element in the DNA of information". [6]The bit is represented by the numerals one and zero, whence its name: digit. Today computers can perform as many as a billion (a million millions) binary calculations/ second, can cipher peoples' actions in all aspects of life into bits and can accumulate the data in limitless databases. Information-storage cost is falling as data-monitoring speed increases.

In parallel developments, people started looking at how computers may be used for mimicking the human mind and interacting with it. Among the key figures in the early days of such cognitive science were J.C.R. Licklider, a researcher in psychoacoustics—seen as the man who planted the seeds of human-computer interaction—and Norbert Wiener, the mathematician who invented cybernetics. This initiated artificial intelligence and personal computing, which pushed the radical idea that computers were usable not only for automation but also for "augmenting human intelligence," as Douglas Engelbart put it. Engelbart is known as the inventor of the mouse but should be credited with much more. On 9 December 1968, Engelbart's team at his SRI Augmentation Center performed what now called "the Mother of All Demos,", where they showed a working prototype of networked personal computing, including the world's first computer mouse, and introducing video conferencing, teleconferencing, email and hypertext.

Engelbart's vision was not taken seriously by everybody at the time. Computers were supposed to be as big as possible. Engelbart suggested they should be as small as possible. Computers were supposed to support special, advanced tasks involving big money, like calculating trajectories of missiles, managing advanced business transactions for large organizations or run nuclear power plants. Engelbart suggested individuals should used them to write letters or chat with colleagues or family members. At the Mother of All Demos, he demonstrated how he edited his wife's shopping list. Personal computing is now an established reality. The Internet is ubiquitous, personal computers are everywhere and we are close to the point when all humans will be using personal computers to augment their intelligence.

As our personal computers are connected on the Internet and used to facilitate interaction, our collective intelligence develops. But so will also the intelligence of the system itself, the artificial intelligence in the systems of machines with which we interact.

More and more information from the analog, real world, is being sampled, digitized and moved into the digital world. There it is stored, communicated and processed in myriads of recombinations, supporting our decision making and making decisions for us.

With the increasingly rapid growth of available data on the Internet, interest is growing in finding ways to create as much value as possible from all this data. It can be about finding new trends in societies of which people have not been aware, or in other ways to find and match bits and pieces of information in different places

[6] Negroponte, Nicholas. **Being digital**, first vintage books edition, Jan 1996.

to produce valuable new knowledge. It is not trivial. It's about finding needles in haystacks, looking for relevant questions to answer, sometimes even looking for answers without knowing the questions.

Data mining addresses these questions. It can be defined as a process of obtaining new knowledge—automatically—by analyzing digital databases (based on binary code), randomly constructed using smart algorithms. An algorithm is an array of commands for the computer to execute, defined actions in a predetermined order. A smart algorithm is an algorithm that does not do the same thing every time but will vary itself in unprecedented ways. It inserts uncertainty into the process in the form of probabilities. Moreover, a smart algorithm "learns from experience (heuristically)" and reduces uncertainty factors over time, thus increasing knowledge, and thus its "wisdom" widens incessantly. One objectives of data mining in the Internet era is to identify hidden characteristics in someone's personality to predict future behavior.

Artificial intelligence is defined as a science that deals with the construction of machines (computers) whose purpose is to perform computing actions and decision-making as an alternative to human intelligence. The strength of smart algorithms is derived from their ability to examine complex situations with many variables, while considering different levels of uncertainty. In this regard we note the important contribution of the English mathematician- theologist Thomas Bayes (1702 – 1761), who developed the basis for the theory that incorporates uncertainty in decision making and problem-solving—a vital element in artificial intelligence.

So, for the first time in human history, the analytical tools of data mining and artificial intelligence enable analysis of situations of uncertainty that have hundreds and thousands of variables. The systems built from these tools are programmed to learn heuristically, constantly reduce uncertainty and thus increase knowledge.

Nowadays, artificially-intelligent algorithms can construct a personal digital identity for every person performing actions on the Internet. Data-mining 'robots' will be able to analyze texts, video and audio contents and transform them into sociological DNA that will describe the individual personality. [7] Constructing the digital identity is a dynamic process updated as long as the person is active on the Web.

2.2 Managing Digital Identities – Developing a Universal Standard

Today, the global knowledge industry invest great resources in the development and improvement of management techniques of digital identities. Digital identity management is developing rapidly and is called "federated identity management"[8].

[7] Lemelshtrich Latar, Noam. **Personal web social DNA and cybernetic decision making**, Hubert burda center for innovative communications, BGU, feb 2004, ICA conference 2004.

[8] Madsen, Paul, **SAML2: The building blocks of federated identity**, Jan 2005, xml.com..

The term "federated identity" refers to various components of users' profiles gathered while they surf on different sites and consolidated into uniform profiles according to a global standard. The term is also used for adoption of standards for the consumer identification process on the various platforms.

A leading global consortium in this field is OASIS[9]: "Organization for the Advancement of Structured Information Standards;" this NPO (non-profit organization) is associated with over a hundred countries and six hundred organizations, including governmental organizations (promoting an electronic government); educational institutions and commercial companies.

An important aspect in the development of a universal standard for creating a digital identity is the Semantic Web and the Socio-Semantic Web (S2W).

The goal of the Semantic Web is to build a universal standard for tagging information on the Internet. Tim Berners-Lee, the founding father of the World Wide Web, who envisions developing the Internet and Cyberspace as a universal medium for information exchange, heads the effort to develop the Semantic Web. The Socio-Semantic Web developed to enable the wide public to tag information on the Web cooperatively (Collaborative Tagging or Folksonomy), to exploit the "collective intelligence" that characterizes the social Internet (Web 2.0) and thus make search engines more efficient.

Developing universal standards for tagging digital identities and their definition is a continuation of the development of the Semantic Web, to make universal tagging of digital identity components more efficient. Currently, the most acclaimed standard for constructing digital identity is called SAML2, "Security Assertions Markup Language 2.0;" it enables consolidation of digital identities of surfers on various platforms and management of those identities; and it allows mobilizing various parts of the surfer's identity definition, defined on different social networks, and merging them into one virtual profile. The standard was successfully assimilated in financial organizations, academic institutions, the American electronic government and more.

The adoption of international standards for defining digital identities is significant vs. the ability to follow up on surfers in any site in cyberspace and carrying out widespread studies on the connection between the users' digital identities and their personaliies, fields of interest and cognitive abilities. Every surfer has a unique, dynamic way of surfing—derived from the person's ability to make decisions, memory and additional cognitive factors—rendered to automatic cognitive diagnosis through the artificial-intelligence algorithms.

Since September 11th 2001 (the collapse of the Twin Towers) and part of facing terrorism, multi-dimensional digital-identity development has accelerated. The American Department of Defense allocated resources for the "Total Information Awareness" (TIA) project through DARPA (Defense Advanced Research Projects Agency, which financed initial development of the Internet), which focused on

[9] http://www.oasis-open.org.

developing smart algorithms for building potential terrorists' digital identities, combined with the person's biometrics (facial structure, pupil structure, fingerprint, walking style, etc.), behavior and actions on the Internet web in education, commerce, tourism, medicine, transportation and housing, as well as all actions pertaining to the government and media. This project's goal is to predict terrorism activities by digital identities. These new research tools will doubtless integrate in the trend of developing digital identity construction.

The Internet was initiated by DARPA for wartime digital communications, looking to design a network that could quickly reroute digital traffic around disabled switching stations. Just as with the Internet, the present defense-driven research on digital identities may generate spinoffs in commercial applications. If digital identities can be constructed for terrorists, simulating their patterns of behavior, enabling others to get the upper hand, the same may go for other target groups—salespeople, shop customers, court juries, parliamentarians or family members.

Just as the Internet created possibilities and stirred pots, so will digital identities. Who will have the right to construct digital identities, mimicking whom? How will such a legal framework be set up and enforced? Many aspects of the process merit discussion. In this essay we note these questions but will not try to answer them.

Afficionados of the '60s TV series "Mission Impossible" may recall the recurring theme, where the "Impossible Mission Force," headed by "Jim Phelps," obtained exceptionally detailed information about some person, enabling the force to train a team member to impersonate him/her (which often including wearing a molded rubber mask). Doing so enabled them to gain access to venues and authority to carry out their operations. On the Internet, this is becoming an increasingly real prospect, for better or worse. Ingredient X, the information enabling the creation of a virtual personality, is becoming readily available on the Internet.

2.3 Digital Identities and Social Networks

Today, many people have started to set up their own digital identities. These digital identities may mimic their real-life identities, or may be new, intentionally different ones.

One of the main Internet uses is activity in social networks. Today, millions of people belong to social networks that answer many needs, social, economical and political. A social network is a group that maintains connection to exchange information in text, video, photos or voice or for social purposes. Every network member must give personal details about themselves, and these are exposed to the other network members or part of them, according to the user's choice.[10]

[10] **Social network sites: definition, history and scholarship**, Dana M. Boyd, school of information, uc Berkeley, and Nicole B. Ellison, Dep. Of telecommunications and information studies, Michigan state university. Journal of computer mediated communications, 13(1), article 11, 2007.

Initial studies show that social-network membership increases the “social capital” of its members.[11] Nevertheless, in this chapter we will not discuss the advantages and disadvantages of social networks as seen by users or society. Our interest lies in the social networks’ demand to provide personal details as a prerequisite to membership.

In the information age ‘friend’ gets a new meaning; traditional ‘friendship’—personal acquaintance, shared experiences, mutual trust, eagerness to invest in friendship maintenance and concern—changes beyond recognition. In the information age “friendship” diminishes often to belonging to a social-network subgroup, most of whose of members are anonymous; "friendship" is essentially measured by the amount of information transferred between the members of the network, and member popularity is measured at times by the number of members interested in receiving information updates from them. Some networks limit their memberships.

Most major social networks have a similar basic architecture. They ask joiners to define their personal profiles and in return offer applications, including the option to upload diverse contents, and provide communication paths between members. In completing the profile, joiners are asked to fill in details about themselves—birth date, sex, marital status, areas of interest, country of origin, religion, ethnic background and even details like prevailing mood. The number of ‘compulsory fields’ varies between networks. Amongst the applications of communication, the social networks facilitate the relay of instant notification (SMS, mails), create discussion groups on certain topics and offer options to participate in synchronous and asynchronous forums.

In the area of content uploads, most large networks let members upload text, video, audio, photos, films, hobbies, feedback on books, places of recreation and more. Many people join different social networks simultaneously, since each fulfills different needs.

In addition to common and well-known social networks whose members join with awareness and free will, latent social networks exist in various organizations. Today, research is done on cross-referencing membership on different networks using smart algorithms, analyzing databases and electronic communication in these networks (social network analysis) to identify hidden social networks and predict the prospect of creating connections between different members on the networks according to their digital identity.[12]

[11] Y. Dubner, Stephen**. Is myspace good for society: a freakonomics Quorum**, Feb. 15 2008, NYT July 2008.

[12] Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Ding, L., Kolari, P., Sheth, A.P., Arpinar, I.B., Joshi, A., and Finin, T. "Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection"**.** In **proceedings of the 15th international conference on World Wide Web.** Edinburgh, Scotland, May 23-26, 2006. www '06. ACM Press, New York, NY, 406-416. DOI=http://doi.acm.org/10.1145/1135777.1135838.

Another field of research focuses on extracting information about network members by cross referencing information between different social networks, either open or latent, to predict conflicts of interest between network members by discovering hidden connections. The researchers Nowell and Klienberg[13] investigated the open network Friend Of A Friend (FOAF), and the bibliography database "Co-authors" DBLP, which includes the names of scientists who participated in studies and mutual publications in computer sciences. The bibliographies network was examined to discover friendship relations and cooperation between different researchers, and identify possible conflicts of interest between potential judges of articles to be reviewed.

An additional example of potential conflicts of interest mentioned in Nowell and Klienberg's article can be found on the social network Linkedin.com, possibly relevant in studying IPOs (Initial Public Offering) or networks like Friendster and Myspace, which contain vast information on social connections to possibly indicate certain inclinations in decision making within different areas.

Some major networks, originally constructed as reservoirs for content to serve the surfers, see their purpose today in providing services, information and products adapted to members' digital identities. In September 2007 the network Myspace informed its shareholders that it intended to undertake data mining, using the profiles and blogs of approximately one hundred million of its members, to direct advertisements and services to them. Thus, this is the start of a screening system that will provide services and information to members according to their digital identity. [14] The declared objective is to improve the membership experience on the network, "to add value to the user experience" (almost a paraphrase on Aldous Huxley, "Brave New World").

Jon Callas, responsible for information security at the company PGP, experts in the Pretty Good Privacy enciphering algorithm (no link to the Personal Genome Project), was astonished by peoples' readiness to reveal personal details on networks: "Always assume that everything you write on the network will be attached to your resume…"[15], whereas the "National Security Agency" (NSA) of the American Ministry of Defense, which_specializes in gathering and analyzing information, is financing a widespread research aiming to "harvest information" from peoples' profiles distributed on social networks.[16]

At the heart of the social network is the graph of social contacts between network members, sometimes extending beyond member groups, for example when

[13] D.L. Nowell., Klienberg, Jan. "The link prediction problem for social networds", **journal of the American society for information and technology**, Vol. 58, issue 7, May 2007. P. 1091-1031, John Wiley and sons, New York.

[14] Abramovitch**,** Gieselle **Myspace has data mining plans**, Sept. 24, 2007. www.dmnews.com.

[15] Marks, Paul. "Pentagon sets its sights on social networking websites", **new scientist**, newscientist.com, Jun 9, 2006

[16] Ibid.

Facebook members tag names of non-members in photos. The who-knows-who graph enables the social-network operator to suggest individual members names of other members they might know but have not added to their 'friends' list. Someone who is a friend of many of your friends might well be your friend. Facebook and other social networks will help you by suggesting "people you might know?"

People will also share information about their states of mind, which things interest them, their whereabouts, and so forth. Just as algorithms may predict who you know algorithms may predict what you are, though you did not tell the network about it yet. If many of your friends share an interest, there is a good chance you will share that interest too.

This turns privacy policies into complicated beings on the social Internet. You might choose not to unveil your secret friends and secret thoughts but there may be enough public information available from you and others for other people to unveil them.

The amount of information that networks accumulate on members is vast, a gold mine for data mining and constructing diverse digital identities on every member. The ability to link members' personal profiles with their data-consumption dynamics on the network will enable the construction of personality-linked databases at an almost scientific level and unprecedented scale. So far, human knowledge in these areas has used classical research models with limited scope, with small, controlled numbers of variables. Performing research that encompasses millions of people—with the utmost accessibility to information and the help of research tools that enable simultaneous examination of hundreds of variables—is a different story.

Social networks create a substantial and dangerous expansion of the digital-identity notion to include complete mapping of surfers' social and professional connections. This mapping will accompany the surfers in all human activities and may become a powerful filter that will limit the information and possibilities presented to them, without them being aware of it.

2.4 Socio-Genetics and Digital Identity

The mind and the body hang together, and science is constantly improving the knowledge about it. We know today that social behavior is linked to genetics. Understanding these connections, and how they work in a social context, is powerful for constructing digital identities and can be valuable for analyzing the body, mind and ecosystem surrounding them: society. So information about people's genetic codes may be as rewarding for constructing digital identities as the information from social networks.

The genetic research and instrumentation for mapping man's genetic code, "gene sequencing", are developing rapidly at leading research institutes and large commercial companies worldwide. Their main objective is to identify genes associated with hereditary diseases and to develop medication based on genetic treatment. Since the completion of the Human Genome Project in 2001, commercial competition has arisen between companies for producing machines that map the genetic code of man. The main research project in this field is the Personal Genome Project (PGP, not to be confused with the PGP Pretty Good Privacy algorithm).[17]

The American company Danaher Motion from New Hampshire, in association with the Church Laboratory for Quantitative Genetic Research from Harvard University School of Medicine, developed the Polonator G.007, based on open code, to identify the genetic mapping of humans at a low cost.[18] George Church, who heads the Harvard laboratory, was an initiator of the Human Genome Project. Church and nine additional scientists (the PGP10) agreed to participate in an experimental group and reveal their personal genetic mapping together, with general information on their lives—intended for cross referencing and connecting between their genetic mapping and their personal attributes, as well as other aspects in their lives.[19] Project PGP aspires to include up to one hundred thousand volunteers in a comprehensive study.

An additional major effort for personal genetic mapping is being done by the company 23andMe, belonging to the wife of Google's founder Sergey Brin. The number 23 in the name represents the number of chromosome pairs in the human body (that is, 'Me and my genetic structure').[20]

The company promises to provide consumers with a possibility to better understand their hereditary traits, their genealogy and likelihood to fall ill, according to the genetic code by analyzing a saliva sample. Via a kit customers receive (Spit Kit), the company analyzes half a million points in their DNA and creates genetic digital identities for them. The customers receive this information by electronic mail, and the digital identities are saved in the company's databases.

The company provides a long list of diseases, as well as of physiological and psychological attributes associated with the human genetic structure on its Internet website; it declares that even though the research is still in its initial stages, it offers to provide its customers with various probabilities concerning fifty seven physiological and mental characteristics derived from the customers' genetic mapping. Amongst these are probabilities for addiction to alcohol and drugs, diseases such as AIDS and cancer, traits such as memory and intelligence level, as well as the tendency to develop various mental diseases. Google recently also decided to finance research at Harvard that will encompass the DNA of approximately one hundred thousand people. The project will begin in the U.S.,

[17] www.personalgenomes.org.

[18] Singer, Emily. "Gene sequencing for the masses", **Technology review**, M.I.T, April 30, 2008. www.personalgenomes.org.

[19] www.personalgenomes.org/pgp10.html. see Singer, Emily, 2008.

[20] www.23andme.com.

Britain, China and Sweden, and its declared objective to create preventive medicine[21].

A research laboratory at Cornell University School of Medicine provides similar services to those of 23andMe. The Cornell laboratory is constructing an Internet website that will enable research participants to access data linking their genetic structure to disease risks, and will let them transfer the data to their personal doctor. According to the laboratory director, Michael Christman, the research in the field is growing rapidly [22] and the laboratory will study approximately ten thousand volunteers in the next two years, including various ethnic groups. Christman calls for the construction of a model for social- ethic reference to studies on this topic, since government and legal authorities are ignoring it and the findings are leaking to the Internet. Meanwhile, other commercial companies, as well as Google, are trying to connect genetic mapping with physiological and psychological aspects, including Decode and Navigenics.

Parallel with the efforts to develop technologies for gene mapping and creating a genetic digital identity, many psychology researchers are working to link the people's genetic structure with their psychological-social characteristics. These studies focus on the connection between the person's genetic structure and negative social behavior (anti-social behavior), such as violence and social aggressiveness, as well as positive social behavior (pro-social behavior), to help others (volunteering, distribution of resources and mental support). The researchers Ariel Knafo and Israel Salomon from the Hebrew University note that the majority of studies so far have focused on the influences of the genetic structure on anti-social activity (about fifty studies); a minority has dealt with the connection between genetics and positive social behavior. Knafo and his group have focused in their study on discovering “the genetic architecture of pro-social behavior”.[23]

So far, research focusing on behavior of identical and almost identical twins has been used to separate hereditary genetic influence and environmental influence, as well as neutralizing the age variable. Knafo and Salomon discovered the connection between a certain gene and altruistic behavior; they recognize that environmental factors exist (parents, family, friends, school, teachers and more), which influence the connection between the genetic structure and social attributes.

Knafo and Salomon, who also reviewed six studies focusing on children and seven studies focusing on adults, found that the studies (except one) indicated a connection between inherited genetics and pro-social behavior, though it is still difficult to identify specific genes associated with attributes in a tested manner.[24]

[21] Singer, Emily. "You've had a genetic test. Now what?" **Technology review.** , M.I.T, June 13, 2008**.**

[22] Ibid.

[23] Knafo, Ariel. And Salomon, Israel. **Genetic and environmental influences on prosocial behavior,** psy. Dep, Hebrew university, June 6, 2008.

[24] Knafo And Salomon, 2008, P.18.

In their book *Society and Psychosis*, Craig Morgan et. al note that studies in the field of Molecular Genetics have revolutionized the field of psychiatric research in searching for links between specific genes and mental diseases: "Recently psychiatric research has been revolutionized by molecular genetics such as the hunt for candidates genes for schizophrenia."[25]

The connection between genes and human traits, and the entry of information-age giants such as Google and leading research centers such as Harvard and Cornell into the field of genetic research, should close the research gaps much faster. The large volume of participants in these studies, the vast databases holding participants' digital identities and data mining peoples' social behavior on the Internet, together with the use of smart algorithms are helping science to begin to predict social behavior, both pro-social and anti-social, according to the genetic mapping of humanity.

2.5 Digital Identities in Virtual Worlds

Virtual worlds were first developed in the field of computer games; millions of people of all ages spend hours playing games in virtual worlds with an imaginary visual multidimensional facet, but close to reality. The Danish company Linden Labs advanced this field with the game SecondLife[26], which created a real interface for the first time interlacing between a virtual game and reality. In this game everyone can create a virtual character, called an Avatar, with which they wander the virtual world and develop an imaginary life while connecting with other Avatars. The Avatar can maintain both a social and a sex life in virtual worlds, participate in discussions and study programs, even create products and sell them for "Linden money," which can be converted to real money on the company's website. Universities like Harvard, MIT, media organizations such as CNN and Reuters, countries like Sweden, as well as economic organizations like IBM, offer real products and services in the SecondLife virtual world.

The use of role play (like psychodrama) is a known psychological tool exercised to remove barriers and discover hidden characteristics in a person's personality. It may be possible to connect the characteristics of a hidden character revealed through the digital Avatar and the person's real life identity, thus enriching the digital identity with additional layers having psychological depth.

Initial studies in this field show a connection between the Avatar character and its attributes, and the person's behavior in reality. It is possible to assume that people operating in virtual worlds are freer of the defense mechanisms guarding their cognitive balance in the real world, and will thereby release information to illuminate more hidden layers in their personality.

So it seems possible that functioning in virtual worlds, being seemingly imaginary, influence peoples' consciousness and actions in reality. Researchers worldwide are

[25] Morgan craig et al. **society and psychosis**, Cambridge university press, march, 2008, p 59.

[26] Secondlife.com.

showing interest in this topic, and the principal assumption is that through a person's activity in virtual worlds you can learn about behavior, personality and opinions in the real world, even influence them.

Bailenson and a team of researchers at Stanford University in the Virtual Human Interaction Laboratory are examining how behavior and appearance of "Avatars" in virtual worlds relate to the use of the computer in reality.[27]

In one study, Bailenson and Yee showed that the representation characteristics of an "Avatar" influence behavior in the virtual world, a phenomenon called "Proteus Effect." In this way, for instance, an Avatar with desirable characteristics acts more confidently in interacting with other Avatars, and tall Avatars act more confidenctly than short Avatars.[28] They also showed that the age chosen for Avatars influences their "behavior" in the virtual world, and reiterates the effect of the game on the behavior in reality.[29]

Another research group, Doron Friedman et. al, developed research robots ("bots") hidden in Avatar characters, which can wander the virtual worlds, communicate with other Avatar characters, question them and transfer the information to researchers in reality.

"We have developed software bots that inhabit the popular on-line social environment SecondLife (SL). Our bots can collect data, engage in simple interaction and carry out simple automated experiment. In this paper we use our bots to study spatial social behavior […] we found that when Avatars were approach by our bot, players tended to respond by moving their Avatars further, indicating the significance of proxemics in SL."[30]

We imagine that such bots—autonomous intelligent agents—will be useful for many purposes and that there will be incentives for their further development. Bots may be designed to develop capabilities of stimulating certain responses from real-life users to harvest information useful for constructing certain digital personalities. Such bots will not only be applied in virtual worlds, but in any social network, including Facebook and LinkedIn. They would be somewhat similar to the search-engine spiders that index Internet Web pages but more intelligent, self-learning and

[27] Fox, J. and Bailenson, J. **virtual exercise in the third person: Identification, physical similarity, and behavioral modeling.** Paper to be presented at the annual conference of the international communication association, Montreal, Quebec, Canada, May 2008..

[28] Yee, N. and Bailenson, J.N. "The Proteus effect: The effect of Transformed self-representation on behavior". **Human communication research,** 33, P. 270-290, 2007.

[29] Yee, N. and Bailenson, J.N. "Walk a mile in digital shoes: The impact of embodied perspective-talking on the reduction of negative stereotyping in immersive virtual environments. **Proceedings of presence 2006: The 9th annual international workshop on presence.** August 24-26, Cleveland, Ohio, USA, 2006.

[30] Friedman, D., steed, A., and Slater, M. Spatial. **social behavior in second Life, Proc. Intelligent Cirtual Agents** LNAI 4722, Pelachaud et al. (eds), p. 252-263, Paris, France, September 2007.

sophisticated. The information they harvest can be used to generate revenues in various ways, similar to the Web spiders.

Today much attention and money is going into virtual worlds and other multi-user on-line games. The multiplayer on-line role-playing game (MMORPG) “World of Warcraft” had 11.5 million monthly subscribers in December 2008 according to its producer Blizzard entertainment, and 62% of the entire massively multiplayer on-line game (MMOG) market in April 2008 according to Woodcock.[31] Game-industry analyst DFC Intelligence reportedly estimated the revenues to around $500 million USD[32].

What began as computer games is thus acquiring great significance in creating the composite digital identity. The virtual worlds are a vast research resource to study hidden facets of the human personality, allowing corroboration of data at a high level to enable the prediction of behavior.

3 Digital Age Journalism

Before asking how AI may be applied in the context of journalism, we need to consider what is happening to the concept of journalism in the digital age. Until now, ‘journalism’ and ‘the media’ have been considered synonyms. Journalism is symbolized by the infrastructure for mass communication and vice versa. “Stop the presses” means “breaking news.”. The organizations controlling the infrastructure for mass communication also controlled the content being broadcast.

3.1 Defining ‘Journalism’ by Relation to People Instead of Relation to Media

This is reflected in the definitions of journalism found in dictionaries, as with the Compact Oxford English Dictionary, published on-line on the Internet through AskOxford.com[33]:

> *journalist • noun a person who writes for newspapers or magazines or prepares news or features to be broadcast on radio or television.*

Ironically, the on-line dictionary does not include the Internet in the list of media. But merely including the Internet would not save the definition. Now everybody can broadcast news over the Internet, but that does not make everybody who does it a journalist.

[31] "MMOG Subscriptions Market Share April 2008". mmogchart.com, Bruce Sterling Woodcock. 2008-04-01. Retrieved 2009-09-03.

[32] “ The top ten money-making MMOs of 2008”, Wagner James Au, GigaOm 2009-02-01, http://gigaom.com/2009/02/01/top-10-money-making-mmos-2008/ Retrieved 2009-09-03

[33] Compact Oxford English Dictionary, published on-line by AskOxford.com. Retrieved Sep 6 2009.

Until now, there have been communication infrastructures for one-to-many communication—media, and one-to-one communication—telephone.

One-to-many communication has been seen as 'the media,' mainly journalism and entertainment, where publishers responsible for broadcasting and consumers have no responsibility—they can choose to receive the broadcast or not. One-to-one communication, mainly telephone, has not been considered 'media' but personal conversations, mediated by an impartial communication infrastructure and telecom service provider. Nobody is responsible for the entire communication, the responsibility is spread between the interacting parties.

With the proliferation of the Internet there is no longer a difference between infrastructure used for one-to-one or one-to-many communication. What's more, it enables many-to-many communication. Web 1.0 spread the one-to-many communication possibility beyond the media. Everybody could publish. Web 2.0 introduces many-to-many communication. Now the crowd can publish together. The new media companies—the ones that do not provide their own content—have no problems with this.

But it brings the old media into an identity crisis. 'The media' implies control over both the medium and the content. The media were the ones who could order "stop the presses" if they thought something important had happened. There has been interaction with the users, but it has been controlled. Media have used phones frequently, but only for media consumers to call in to be a part of the broadcast. It has enabled media companies to include the audience in the broadcast, without losing control.

Now that has all changed with the Internet. In trying to preserve their practices and identities, media companies tend to lean toward holding on to dated one-to-many media technologies. Their business models, based on controlling the medium and the content, have been difficult to move to the Internet. In the cases where new business models for ads have succeeded, such as Google, eBay or Craigslist, the brokering of ads is not integrated with the practice journalism.

The essence of journalism remains to be described in 'principles of journalism, such as those suggested by the Pew Research Center's Project for Excellence in Journalism (PEJ) and the Committee of Concerned Journalists[34]:

1. Journalism's first obligation is to the truth;
2. Its first loyalty is to the citizens;
3. Its essence is a discipline of verification;
4. Its practitioners must maintain independence from those they cover;
5. It must serve as an independent monitor of power;

[34] PEJ and CCJ Principles of Journalism, published 1997. Available at http://www.journalism.org/resources/principles Retrieved 6 Sep 2009.

6. It must provide a forum for public criticism and compromise;
7. It must strive to make the significant interesting and relevant;
8. It must keep the news comprehensive and proportional;
9. Its practitioners must be allowed to practice their personal conscience.

These principles remain, even when we no longer know what "the media" are.

To conclude, it can be constructive to look for a new short definition of 'journalism', separating it from 'the media'. Such a definition should connect to the principles of journalism, and be based on the relation between journalism and its audience, rather than on its relation to the medium it uses for communicating with the audience (which is what is causing the confusion today). For example, as in the following suggestion[35]:

Journalism is the production of news stories, bringing public attention to issues that interest the public. Journalism gets its mandate from the audience.

It must act in the interest of its audience. It is not performed on behalf of its sources or its advertisers. When attention work is done in the interest of the sources, it is PR, not journalism.

Journalism need not always fund itself by generating and brokering attention, i.e. attention work. It might also fund itself by generating and brokering knowledge. In this case users will be paying for knowledge. Indeed, this has already been a part of the journalism business model—people have over time paid for a commercial-free cable news subscription and so forth. Some journalistic products have been funded mainly by payments from subscribers, such as newsletters. With a knowledge business model, journalism approaches the work done by analysts, who also keep themselves going by selling knowledge rather than attention.

This said, journalism's role as the agenda-setter of public debate (as described by McCombs and Shaw in their agenda-setting theory)[36] depends on journalism's ability to focus public attention on issues that interest the public. This is for many a *raison-d'etre* for journalism, which requires a business model that incentivizes attention. The knowledge business models for journalism, such as newsletters, do not necessarily incentivize broad public attention and may in fact dis-incentivize it. Who will pay a substantial price for a newsletter containing information that already is known to the public, and probably available on the Internet?

[35] Nordfors, David 2009, "Innovation Journalism, Attention Work and the Innovation Economy. A Review of the Innovation Journalism Initiative 2003-2009", Innovation Journalism Vol 6 No 1, http://www.innovationjournalism.org/archive/injo-6-1.pdf , Retrieved Sep 9 2009.

[36] McCombs, M.E., and D.L. Shaw. (1972) The Agenda-Setting Function of Mass Media. Public Opinion Quarterly, Vol. 36 p.176-187

At the end of the day, the central question for the continued survival of journalism as an independent business is not how to adapt to the Internet, but what are the business models for assuring the principles of journalism in the innovation economy.

3.2 Augmented Human Intelligence, Collective Intelligence, AI and Journalism

As Information Technology and the Internet proliferate, IT-assisted intelligence is growing. This is the 'augmented intelligence' of individuals that Doug Engelbart addressed with personal computer systems, the 'collective intelligence' enabled by putting together personal computers in networks, enabling groups of people to act intelligently together. It is also about Artificial Intelligence (AI), the ability of machines to act intelligently without human assistance.

These IT-assisted intelligences blend. Collective intelligence and AI take part in improving augmented human intelligence and vice versa. We are improving the machine and the machine is improving us. We are becoming the machine and the machine is becoming us.

Journalism, described by its principles, is also becoming a part of the machine, just as the machine is becoming a part of journalism. Algorithms are being developed for rating news, based on mixing systems for aggregating opinions of the crowd (collective intelligence) and smart algorithms for contextual analysis (AI).

Journalism's role is to focus public attention on stories that interest the public. For journalism to remain a meaningful job for its practitioners, it should also empower the audience. Seen from this angle, it is relevant to discuss how journalism interacts with augmented intelligence of individuals, the collective intelligence of society, and the artificial intelligence of machines. Ideally, journalism raises intelligence—empowering the audience—as it improves itself from the higher intelligence of the system surrounding it, i.e. the audience and the machines.

Just as it makes sense to look at how journalism can address digital identities in its creation of stories, its interaction with the audience and its business models, it makes sense to look at the digital identity and DNA of journalism itself.

4 AI AND JOURNALISM CONTENT ANALYSIS

4.1 Establishing the DNA of Journalistic Content

From the early nineties major interdisciplinary research efforts have been invested in developing efficient ways to automatically retrieve information and knowledge from multi-model journalistic content. The main objective of this research is to let consumers find information they seek quickly and effectively. Today, the major search engines like Google, Yahoo and others yield millions of links to any request and cannot answer consumer requests expressed by simple keywords. The community of researchers involved in this multimedia information retrieval research (MIR) covers Human Computer Interaction (HCI), Information Theory (IT), Statistics, Pattern Recognition, Psychology and recently, the Social Sciences. Recent papers in these separate fields offer citations and borrow research methods and tools from the other fields.[37]

Visual content analysis was performed using meta-data or by using text annotations added manually and more recently by data mining audio tracks that accompanied the visual material. But these methods are insufficient in today's multi-media vast journalism content being created and deposited in digital warehouses. Automatic, computer-based multi-media content analysis is the only answer if humanity is to benefit from the vast digital data deposits created daily. "What good is all the knowledge in the world if one can't find anything?" [38]

The substantial multi-disciplinary research into information retrieval from journalistic content is done from the perspective of the consumer-initiated search for information and knowledge. Our objective is to study the implications of the reserve prospective, analyzing the significance of a new journalistic phenomenon where the content automatically searches the consumers based on their digital identities discussed in the previous section. But first we describe current research on automatic knowledge retrieval from journalism multimedia content.

4.2 Journalism Multimedia Content Analysis

Content-based information retrieval covers several research communities: Content Based Image Retrieval (CBIR), Content Based Visual Information Retrieval (CBVIR), Multimedia Retrieval (MIR), Automatic Image Annotation and Retrieval (AIAR), Cross Media Relevance Models (CMRM), Automatic Text Analysis, Audio analysis groups and others. As will be shown, most research tools used by the different communities aim at dividing content into small content digital units, analyzing them, tagging these sub-content units and then carrying an integrative analysis to conceptualize the entire content meaningfully for the consumers. Some researchers convert the visual content into mathematical formulations which can then be subjected to analysis employing artificial intelligence algorithms.[39]

[37] Ritendea Datta, Dhiraj Joshi, Jia Li and James Z. Wang, "Image Retrieval: Ideas, Influences and Trends of the New Age", ACM Computing Surveys, Vol 40, No 2, article 5 pp. 1-60 2008

[38] Michael S. Lew, Nicu Sebe, Chabane Djeraba and Ramesh Jain, "Content Based Multimedia Information Retrieval: State of the Art and Challenges", ACM Transactions on Multimedia Computing Communications and Applications, Feb 2006.

[39] J. Jeon, V. Laverenko and R. Mammatha, "Automatic Image Annotation and Retrieval Using Cross Media Relevance Models", ACM 1-58113-646, 3/03/2007

4.3 Content Based Image Retrieval (CBIR)

Research on retrieving images from visual content has seen significant growth in the past decade.[40] Datta *et al*, surveyed over 300 key theories and empirical research contributions in their paper. "Computer vision, machine learning, human computer interaction, statistics, psychology… are becoming a part of CBIR"[41] CBIR is defined as "any technology that in principle help organize digital picture archives by their visual content."[42] Until recently the search of visual images was based on analyzing texts added manually to describe images or manual image annotation.[43]

The major objective of the CBIR community is to reduce what is termed "semantic gap" between the human manual coding and the automatic machine coding of the same visual content.[44]

A vast amount of voluntary human annotation of visual content can be found in the popular website of YouTube and Flickr and can serve as reference data for calibrating the automatic tools. These social network sites allow for the 'wisdom of the crowd' to assign tagging to visual content. Face-recognition techniques and medical applications have contributed greatly to the advance of CBIR.

The primary method used in the search for image retrieval or automatically conceptualizing visual content is dividing the visual frames into smaller sections/regions termed as "blobs." This is achieved by using statistical tools such as clustering. Each blob is annotated with text. The visual image is described by employing categories such as color, texture, shapes, and structures. Statistical theories are used to associate words with image regions that are then compared with human manual annotations of similar images.[45] Attempts are made to describe

[40] Note 1, Supr

[40] Wang, J.Z. Boujemah, N. Del Bimbo, A Geman, D. Hauptmann, A, and Tesic J., "Diversity in Multimedia Information Retrieval Research", MIR Workshop, ACM Multimedia 2006.

[40] Note 1 Supra, Page 2

[40] Note 3 Supra

[40] Smeulders, AW Worring, M Santini, S Gupta and Jain R, "Content Based Image Retrieval at the End of a

[41] Wang, J.Z. Boujemah, N. Del Bimbo, A Geman, D. Hauptmann, A, and Tesic J., "Diversity in Multimedia Information Retrieval Research", MIR Workshop, ACM Multimedia 2006.

[42] Note 1 Supra, Page 2

[43] Note 3 Supra

[44] Smeulders, AW Worring, M Santini, S Gupta and Jain R, "Content Based Image Retrieval at the End of the Early Years", IEEE Trans., Pattern Analysis and Machine Intelligence, 22, 12, 1349-1380, 2000

[45] Note 3 Supra

images using "vocabulary of blobs" as proposed by Duygulu *et al*.[46] Jeon *et al* proposed a method for using "training set of annotated images for … cross media relevance model for images."[47]

Today the CBIR is being applied to domains such as "family album management, botany, astronomy… the interpretation of art and cultural images… to massive picture archives and travel photography"[48] and automatic indexing of pictures in real time automatic text annotation.[49]

CBIR researchers are developing mathematical descriptions of images defined as 'signatures.' The signatures describe an image in mathematical formulations to let researchers measure content similarities between image frames. Statistical methods such as clustering and classification are used to form image signatures that will allow automatic similarity measurements by machines. Images are segmented by features such as color, texture differences, shapes and other salient points.

Digitizing an image allows the researcher to transform the image into a large set of pixels (picture elements), each pixel having a uniform color. A pixel is defined as the smallest item of information in an image, usually a square. Dividing the entire image into small units of uniform color lets the researcher transform the entire image into mathematical formulations. This is also the basis for image compression for efficient storage and transmission. Grasslands, skies, water, represent regions of pixels with similar colors.

4.4 Video Information Retrieval

In video retrieval, researchers have attempted to develop automatic retrieval methods that do not rely on human subjective analysis. This required developing techniques that identify thresholds between color histograms corresponding to consecutive video frames.[50] A search engine "ImageSpace" was developed where users could direct queries for multiple visual objects, such as sky, trees, water etc. These tools were used for several video searches including automatic detection of pornographic content.[51]

[46] P. Duygulu, K Barnard, N. de Freitas and D. Forsyth, "Object Recognition as Machine Translation: Learning a Lexicon for Fixed Image Vocabulary", Seventh European Conference on Computer Vision, pages 97-112, 2002.

[47] Note 3 Supra

[48] Note 1 Supra, Page 15

[49] Li, J and Wang J.Z. "Real Time Computerized Annotation of Pictures" ACM Multimedia 2006

[50] Flickner, M Sawhney, H Niblack et al, "Query by Image and Video Content: The QBIC System", IEEE Computer, Sept; 23-32, 1995

[51] Note 2 Supra

4.5 Fusion of Research Methods for Content Retrieval

Most of the experimental work, until recently, was performed in each medium content separately. At present, the CBIR community recognizes that integrating research methods is needed to reduce errors and lower the 'semantic gap' between the computer and human analysis. These new research efforts are labeled 'multimodel fusion.' An example is mining video data pre-annotated with text or a soundtrack analyzed in conjunction with visual 'signature.' This integration or fusion should yield better content conceptualization. Fusion analysis of video content allows for important video applications, such as analyzing scene changes and story segmentation.[52]

4.6 Text to Pictures – the Reverse Direction

Recent attempts have been made to attach visual representation for a given text or story[53]. This reverse action can be useful for illustrating journalism news stories, as frequently done today, by adding animations to explain events for which no video content is available. These efforts, still in their infancy, pose a great challenge. It is too early to judge the scientific contribution of this new direction, but it may lead to a better understanding of automatic image annotation.

4.7 Human Centered Content Analysis

It has been long recognized that human satisfaction with search for knowledge and information in multimedia content involves several dimensions. The search process involves a mixture of rational as well as emotional dimensions. The consumer search takes place in a certain context and emotional state and identical search results may be viewed differently by the same person as based on his or her emotional state at the time of the search. A person's background, education and values, affect his or her satisfaction with the search results.

The semantic gap discussed so far is just one dimension with which to evaluate search results. Another important dimension is to study the emotions that a certain piece of content evokes in people. Datta *et al* refer to this dimension as aesthetics; "Aesthetics is the kind of emotion a picture arouses in people. Emotions are subjective reactions and should be measured as such. This is referred to at times as affective computing… which focuses on understanding the user's emotional state which affects his satisfaction with the information retrieval."[54]

[52] Zhai y., Yilmaz A. and Shah M. "Story Segmentation in News Videos Using Visual and Textual Cues" in ACM Multimedia, 2005

[53] Barnard K, Duygulu, P Forssyth et al, "Machine Words and Pictures" J of Machine Learning Research, 3, 1107-1135, 2003

[54] Note 1 Supra, page 46

CBIR researchers realize that integrating human feedback and involvement in the automatic multi-model content analysis is crucial in reducing errors and increasing user satisfaction. This new direction in research is called 'human centered computing.' "By human centered, we mean systems which consider the behavior and the needs of the human user. Other researchers call the human centered research affective computing. Affective computing seeks to provide better interaction with the user by understanding the user's emotional state, and responding in a way which influences or takes into account the user's emotions."[55] Some researchers attempt to define images according to emotional categories. Salway *et al* developed a way to extract character emotions from films based on a model that links character emotions to events in their environment.[56]

4.8 The Aesthetic Gap

An attempt to measure emotions vis-a-vis content analysis beyond the "semantic gap" led Datta *et al* to define the 'aesthetic gap.' "The aesthetic gap is the lack of coincidence between the information that one can extract from low-level visual data (i.e. pixels in digital images) and the interpretation of the emotions that the visual data may arouse in a particular user in a given situation."[57]

Realization that human inputs are required to reduce 'semantic gaps' and 'aesthetic gaps' to achieve meaningful automatic journalistic multimedia content leads to our estimation that the integration between the automatic computer content analysis and the personal digital identities will, over time, allow for a meaningful, personally adjusted content. As will be shown in section four, automatic AI engines already used by governments for delivering services to their citizens may be used to automatically target journalistic content per the consumer personality. The comprehensive, automated content analysis will also consider the context of the content consumption. There may be numerous content 'units' conceptualizations per a given visual content, as we know from traditional mass-media research. People interpret texts and visual messages differently based on their background, political inclinations and values. Only this time these interpretations will be done automatically by computers.

4.9 The DNA of Music.

A unique contribution to CBIR comes from an unrelated field, a Spanish high-tech company, Polyphonic, a Barcelona based music intelligence solutions company[58]. Polyphonic employed a mathematical model to discover the DNA of successful music. The company's goal was purely commercial, to reveal to music composers

[55] **Note 2 Supra**

[56] **Salway A and Graham M. "Extracting Information about Emotions in Films"**, Proceedings of the ACM International Conference on Multimedia, Berkley, USA, Nov. 2003, p. 299-302

[57] Note 1 Supra, page 46

[58] Polyphonic.com

the secret of successful music composition, or what we term the DNA of successful music. In Polyphonic's case, consumers are viewed as one whole homogenous personality composed of six billion people. The data-mining algorithm developed by Polyphonic is called "HIT Song Science (HSS)" and its purpose is to predict the success or failure of music. In CBIR terms, the Polyphonic algorithm can be used to retrieve music from large data warehouses that fit the mathematical description of a successful song 'signature.'

Polyphonic produced a vast digital library of one million music tracks. The company studied three and a half million successful music compositions and created mathematical signatures found to be present frequently in what people, over the years, determined to be popular music. According to Polyphonic scientists, all successful (popular) music has similar patterns and falls into similar clusters. If you wish your song to be successful, it must fit these patterns. The mathematical formulae may change according to different music genres such as: classic, rock, punk, country. But successful compositions in each genre have a similar mathematical signature.

The algorithm analyzes the music based on over sixty elements of the song including its melody, harmony, tempo, beat, pitch, octave, rhythm, brilliance and cord progression, and compares it to over three and a half million past commercial hits. The program organizes the songs into clusters and ranks them "curiously, the clusters of songs do not necessarily contain songs that sound the same to the human ear, but from a mathematical perspective they share similarities."[59]

The program has already successfully predicted the success or failure of songs. Record companies are now becoming Polyphonic customers. It is beyond the scope of this paper to discuss the possible implications of the Polyphonic algorithm to the future of music composition. Some predict it could damage music creativity as composers, instead of trying to be artistically creative, will try to create music that would cater to popular tastes as expressed by the algorithm. But the concept of converting audio compositions to mathematical formulations, clustering the songs along consumer satisfaction and predicting human behavior based on the song signatures is relevant to our discussion of the CBIR. A music piece can be compared to other artistic creations like painting, sculpture and video clips.

Polyphonic's methods are similar to the academic research methods described in relation to the CBIR. A possible next step for Polyphonic could be to divide consumer personalities into clusters per their digital identities, then refine the formulae to specific music signatures per the cognitive and other attributes of the consumers. Polyphonic offers significant validation to the relevance of employing AI to create mathematical formulation of content for the journalism content, as a real-world market validation.

[59] Charlie Devereux, CNN.com, Europe, 17.3.2008

4.10 The DNA of Literature

Way before the invention of computers (but after the invention of the Boolian Logic and Bayes Theorems that laid the mathematical foundations of modern computers and algorithms), in the mid nineteenth century, a French writer, George Polti, (b.1868) analyzed the elements of successful literature, or in other words, its DNA. Polti composed a list of thirty-six dramatic situations that appear in good drama. These dramatic situations include: prayer to the supernatural, crime pursued by vengeance, loss of a loved one, recovery of a lost one, disaster, remorse, revolt against a tyrant, enigma, and others. Polti's list remains popular until today and writers frequently use it in developing stories. Terry Rusio, the script writer for Shrek, said he referred to Polti's list to resolve a situation in the film's plot. To create his list, Polti analyzed classical Greek texts and French literature. Polti's book "Thirty Six Dramatic Situations" can still be purchased at Amazon.com. Polti's analysis of the DNA of drama was followed by other writers.

These attempts to discover good story elements, good drama, were not written from an information-retrieval perspective but provide the literary 'blobs' that will let CBIR researchers dissect, in our case, a journalism story along its content elements. These content elements can be found in texts as well as video scenes and these literary and music 'blobs' may receive mathematical formulations that will allow computer-based analysis. It can be expected that one will be able to retrieve multimedia content based on Polti's thirty six situations for comparative analysis or for marketing news stories to consumers based on their digital identities.

5 JOURNALISM CONTENT AND CONSUMER ENGAGEMENT

5.1 The Concept of Media Engagement

The economic engine that until now has driven journalistic activity in the non-public service media is advertising based. Journalism companies, regardless of their media platform (paper, video, or audio) sell consumer attention to advertising agencies. Though inaccurate, rating was the key measuring tool until the advance of the Internet. All the pre-internet rating techniques cannot measure the amount of real attention given by consumers to specific journalistic content. New media platforms, with the advance of the interactive nature of new media, make the competition for consumer attention fierce and complex. The journalism industry now invests great resources to develop new ways to measure consumer attention along multiple parameters, including the consumer cognitive and behavioral profiles and the context parameters of the content. The interactive nature of the new media platforms begins to allow for scientific measurement of consumer attention.

In this new battle for consumer attention the concept of 'engagement,' a relatively new term, is being used to describe the new' relations' between consumers and journalistic content. The Advertising Research Council (ARC) has devised the following definition of media engagement: "Engagement is turning on a prospect to a brand idea enhanced by the surrounding context… the working definition proposed by ARC encapsulates the ultimate objective of linking positive effects towards a brand with brand advertising within the environment of the program content."[60]

Notice that context within which content is delivered is becoming of prime importance. Kilger *et al* identified three important mechanisms involved in enhancing consumer engagement in a journalistic content[61]:

- **Cognitive** (relevance of the program and advertisement to the consumer)
- **Emotional** (the extent to which one likes the content and advertising)
- **Behavioral** (paying attention to the program and advertising content)

The main hypothesis is that the more engaged the consumers are the more they will spend on the advertised product. This recognition by the advertising world, that consumer engagement in journalistic content involves consumer cognition, emotional profile and behavior, provides significant relevance to the computer based information retrieval as applied to content analysis discussed above.

Research conducted by Kilger *et al* about the relationship between media engagement and product purchase likelihood reveals that as engagement measures increased so did the mean likelihood of products advertised in the media to be purchased. Three media platforms were studied—television, Internet and printed magazines. All three platforms exhibited similar findings. "Internet and magazines exhibited very close response curves, while TV followed a similar path but slightly lower mean of purchase likelihood."[62]

The personal parameters Kilger *et al* examined were those traditionally used in social-science research: gender, age, education, income, race, material status. Age, income and race did matter. In the TV and Internet, people with lower education expressed higher levels of trust in the media and older people reported lower engagement. The finding that personal attributes do affect media engagement, as can be expected, is of great relevance regarding the digital identities discussed in the first section. Digital identifies are valuable to advertisers, who will not hesitate to take advantage of them once available on a large scale and accessible automatically. The road to influence journalistic content in the direction of higher consumer engagement is short. The Kilger *et al* considered a limited number of

[60] Kilger Max, and Romer E, "Do Measures of Media Engagement Correlate with Product Purchase Likelihood?", Journal of Advertising Research, p 313, Sept. 2007.

[61] Ibid.

[62] Ibid.

personal parameters as a larger number was "too many to fill within the space constraints of this article."[63]

The Internet not only offers possibilities to measure and broker consumer attention and engagement, but also consumer interaction. A pay-per-click model does this, as advertisers will pay not for being visible but for consumer clicks, an action. This can be taken further. For example, a click on an ad will usually lead to a sales site and may result in further interaction between the consumer and the vendor, including a purchase. So ads, therefore journalistic content, could in principle be paid by finders' fees. This could, however, introduce business incentives for journalists that might jeopardize journalistic principles.

To convert the content-engagement/ product-purchase relations into a science requires the analysis of many variables including contextual ones, and requires automation and the introduction of artificial intelligence: "Excelling during an era of frugality in high expectations requires digital marketers to be accountable for every dollar…The ROI focus will force agencies… to improve effectiveness and we see increased independence on automation… recent shifts [in the liberal direction] in user privacy perceptions have created a window for marketers to use artificial intelligence to run efficient campaigns."[64]

We address how AI is going to be employed in the next section.The ultimate goal of engagement as perceived by the advertising industry is to target advertisements to consumers based on contextual and personal parameters as listed by the Kilger group: cognition, emotions and behavior. This today is being done and researched in the new media channels and is termed by the academic researchers, journalists and the advertising industry 'Behavioral Targeting.'

5.2 Behavioral Targeting and Journalistic Content

The advance in the Internet and Web 2.0 interactivity, characterized by consumers becoming content creators and providers, opened new frontiers for targeting advertisements directly to specific consumers based on their interactive behavior. As of the late 1990s a new marketing field gained significant academic and industry attention: behavioral targeting. "Behavioral Targeting is the ability to deliver advertisements to consumers based upon their behavior while viewing web pages, shopping on-line for products and services, typing keywords into search engines or combinations of all three…"[65]

Over thirty major Internet companies are involved in behavioral targeting, including Google, Microsoft and Yahoo. M. Kassner[66] surveyed Google's

[63] Ibid.

[64] Ibid.

[65] Debra Aho Williamson, White Paper on Behavioral Targeting, Wall Street Journal and e Marketer, May 11, 2005

[66] Kassner M.,"Google Quitely Starts behavioral targeting" ZDNetAsia, April 21 09.

extensive use of behavioral targeting. Google confirms this in its official website. Google operates in two separate systems, Adwords and AdSense. Adwords targets its advertisements based on the search subject matter by identifying keywords in the search. AdSense targets ads based on website content the consumer views "for example if you visit a gardening site, advertisements on that site may be related to gardening."[67] AdSense was extended to searches of annotated images and to annotated videos in YouTube. According to Kassner, "Google is also trying to present relevant advertisements in the Gmail application…by scanning every Gmail message for spam… and sending ads based on the keywords… the whole process is automated and involves no human matching ads to the Gmail content."[68] Google's rationale is that by making ads more relevant to customers it brings them more value. So far, AdSense and Adwords, in all their applications, are still based on text analysis. Once image and video content are analyzed and annotated automatically, as described earlier, behavioral targeting will likely be applied to all journalistic content.

5.3 Behavioral Targeting in Social Networks

Social networks characterized by voluntary profiling by members uploading vast amounts of personal data in texts, pictures and videos are becoming fertile grounds for behavioral targeting. Social network members' profiles include lists of their friends, hobbies, demographics and other areas of interests. Behavioral targeting activity is rapidly growing in social networks. Advertising startups are beginning to develop behavioral targeting technologies especially developed for social networks. Stefanie Olson[69] describes one such example the 33Across.com. The New York based company 33Across algorithms can follow consumer behavior patterns in social networks and are able to identify "sociograms" among the members and can identify for advertisers the more influential members and the "viral propagators" by studying the message dynamics. Universal Pictures is employing 33Across to study how people share their studio trailers or content with their friends.[70] Other companies, which started to employ behavioral targeting on social networks for marketing advertisements include Reverence Science and Tacoda Systems which was bought by AOL and is now a full subsidiary of AOL. Yahoo launched its own activity in social networks called SmartAds, which combines behavioral information with demographic data for targeting ads. Behavioral targeting ads spending is projected to be around one billion dollars during 2010 and grow to 3.8 billion by 2011.[71]

[67] Ibid.

[68] Ibid.

[69] Stefanie Olsen, "33Across: The Next Generation of Behavioral ad Targeting", news.cnet.com, June 23, 2008.

[70] Ibid.

[71] Mills Eleanor; "AOL buys ads from Tocada" ZDNetAsia, July 25, 07

Behavioral targeting raises serious issues regarding the issue of privacy, discussed extensively in academic literature and political circles. The issue of privacy vis-a-vis consumer profiling is beyond the scope of this paper. Tim Berners-Lee, credited with inventing the World Wide Web, spoke before the U.K. parliament on the issue of privacy and the Internet. He said that he came to "raise awareness to the technical, legal and ethical implications of the interception and profiling by ISPs in collaboration with behavioral targeting companies."[72] Berners Lee went on to say that "it is very important that when you click , you click without a thought that a third party knows what we are clicking on… I have come here to defend the Internet as a medium."[73]

But a growing number of surveys done by TRUSTe (a company specializing in privacy matters) shows that the public "certainly show a willingness… to submit to monitoring and enhanced content delivery."[74] This is a remarkable finding that should be followed.

5.3.1 Project "Smart Push"

Davitz of SRI applies machine-learning techniques to study communications in social networks as part of a multimillion dollar project funded by the Defense Advance Research Project Agency (DARPA) of the US Defense. DARPA funded the original research, which led to the development of the Internet. Davitz's objective was to "automatically monitor people's interest and influence in military communities… to identify the influencers… and then to ensure that they see relevant information in news feed to that topic."[75]

Davitz calls this targeting of news according to members' interest profiles "Smart Push." According to Olson, "SRI is looking at commercial applications for it not related to advertising… you can already learn more about people from MySpace and Facebook."[76]

When a powerful research institute like SRI promotes concepts like "Smart Push," it can be expected that news media, where 'rating is king,' will adjust journalistic content in all its media platforms to fit consumers' digital profiles. In the following section we describe how this may be done by using an AI engine, already in government use, to filter or 'webline' services based on digital identities.

[72] Frank Watson "Behavioral Targeting: Profiling or Projecting User Experience" Search Engine Watch, Mar 13. 09.

[73] Ibid

[74] Ibid

[75] Note 33 Supra

[76] Ibids

6 AI AND BEHAVIORAL TARGETING ENGINE AND DIGITAL IDENTITIES

6.1 Automation, Cybernetics and Social Development

Cybernetic theory started to develop in the '40s during the Second World War, as part of the United States' war effort. A group of scientists from various fields, headed by Norbert Wiener, an MIT mathematics professor, laid the foundations for transferring decision-making processes from human factors to 'machines'—the computers. In his book 'Cybernetics,' Wiener writes:[77] "the entire sequence of operations be laid out on the machine itself so that there should be no human intervention from the time the data were entered until the final results would be taken off: and that all logical decisions necessary for this should be built into the machine itself."

Cybernetics recognizes the importance of control and decision-making processes and is aware of their influence on the ability of biological, social and mechanical systems to exist. An organization's survival and its ability to face environmental changes, expected and unexpected, as well as finding a new balance in a changing reality, depend on its capacity to transfer information and process it quickly, effectively and precisely. According to Wiener and his group, the ability to quickly and accurately process a lot of information necessitates a transition to automated decision making processes by computers, without involvement of the human factor.

The vision illustrated in Wiener's book is the basis for the transition to the electronic government and automated decision-making processes, needed in light of the information overload.

Twenty years later, a theory of social development was developed based on Cybernetics, which emphasized the need to automate decision-making processes in social systems. According to this theory, as an organization develops, its complexity grows as it accumulates more information. The organization must develop tools to deal with the increased complexity, filter information, manage it and derive knowledge from it. Developers of the theory were the Czech scientist Arab Ogly and Ernest Kolman, who wrote about it in the Soviet literature (the regime in the Soviet Union considered Cybernetics effective for social control and surveillance, and did not observe that authentic transfer of information is essential for automated decision-making processes, so the cybernetic model cannot work under censorship and without freedom of expression).

"Species which can not achieve a stable dynamically equilibrated interrelation with their environments , retrogress. Those systems progress which can maintain a

[77] Wiener, Norbert . **Cybernetics,** John Wiley and sons, The technology press, N.Y, N,Y, 1948.

homeostatic stability in relation with their environment. At certain levels of complexity of organization automatic feedback systems must evolve if homeostasis is to be maintained. […] automation is therefore a universal law of development"[78]

The social development theory set by Ogly and his associates, as well as the cybernetic principles established by Norbert Wiener and his interdisciplinary team, are the conceptual foundations for transition to the electronic government and the application of automated decision- making processes.

The information age illustrates the processes described in the Cybernetic social-development theory: information created by society doubles every few months; the digitization processes of all peoples' actions on the various media channels accumulate in immense databases; and inasmuch as accessibility to information is more effective, the organization will have more solutions to function especially during unexpected environmental, social or economic changes.

6.2 Digital Identities in Decision Making Processes in E-Government

An additional aspect concerning the digital identity is the rapid transition to an electronic government where communication between citizen and government is electronic, through the Internet. Many countries have passed laws that compel change to a "paperless government" in which all services are provided through the web. In October 1998 the Government Paperwork Elimination Act came into effect in the United States; it induces all the authorities to develop or purchase information technologies as a substitute for paper within five years.[79]

Transition of organizations to Cyberspace and the creation of huge amounts of digital data calls for a shift to new models of decision making in many areas previously governed by the human intelligence knowledge and intuition, for all their advantages and disadvantages.

Decision making in economics and psychology have long been studied, such as in the comprehensive study by Daniel Kahneman (who received a Nobel Prize for it in 2000) and Amos Tversky, who published the now famous work on "Judgment under uncertainty heuristic and biases".[80] Kahneman, Tversky and others proved that people do not make decisions rationally. Research showed that people are influenced variously by the level of uncertainty of the decision-making process directly affecting the results. Personality traits play a significant part in the process. The transition to automated decision-making processes requires new and complex

[78] Ford, Y.Y. "soviet cybernetics and international development", in: Dechert, Charles (ed.) **The social impact of cybernetics, A clarion brok**, simon and Schuster, P. 172, N.Y, N.Y, 1966, p 172, 175.

[79] http://cio.gov/documents/paperwork_elimination_act.html.

[80] Tversky, Amos., Kahneman, Daniel. "Judgment under uncertainty heuristic and biases", **science, new series**, Vol. 185, No. 4157, pp. 1124-1131, sept 27, 1974.

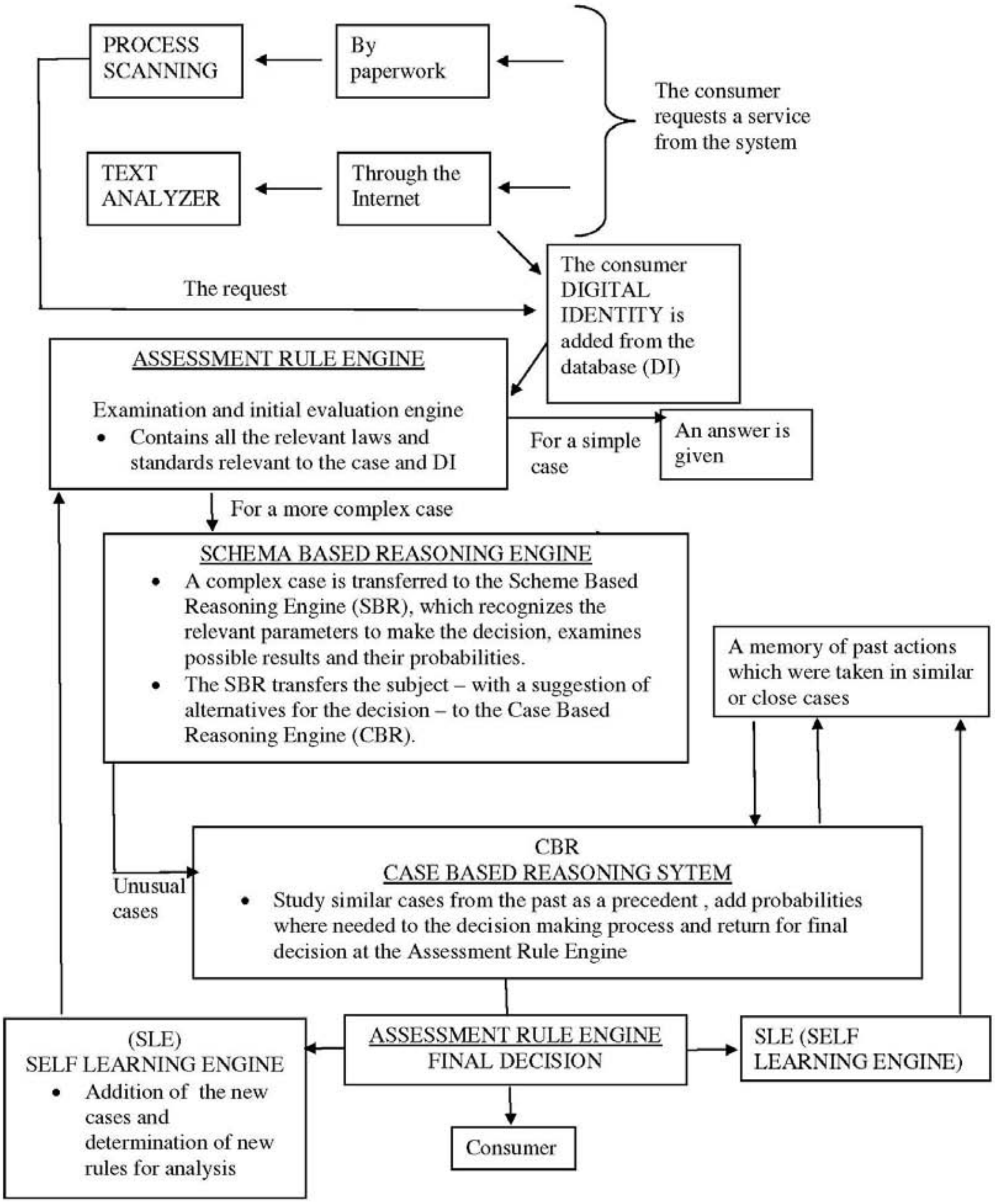
models that will allow replacing human intuitive thought processes with fully automated, AI-based decision-making procedures.

The automated AI decision-making models adopted by several governments comprise elements known as ‘engines.’ The ‘work flow’ of automated decision making is designed to simulate the human mental decision-making process. The advantage of the AI 'decision engine' is that it forgets nothing and can scan large databases in a split second, to calculate the levels of uncertainty including many components in the decision process, to learn from experience and past mistakes, as well as to be updated dynamically. These qualities are meant to prevent the collapse of the system due to information overload, and significantly improve decision making.

The following flow chart describes one such AI automated decision making process already adopted:[81]

[81] Hou Wai Chun, Andy**. Using AI for e-government automatic assessment of immigration applications forms**, university of Hong-Kong, dep. Of science, 2007, association for the advancement of AI. www.aaai.org.

Decision Making Process – The Artificial Intelligence Model


PROCESS SCANNING
By paperwork
TEXT ANALYZER
Through the Internet
The consumer requests a service from the system
The request
The consumer DIGITAL IDENTITY is added from the database (DI)
ASSESSMENT RULE ENGINE
For a simple case
An answer is given
For a more complex case
SCHEMA BASED REASONING ENGINE
A memory of past actions which were taken in similar or close cases
Unusual cases
CBR
CASE BASED REASONING SYTEM
ASSESSMENT RULE ENGINE
FINAL DECISION
(SLE)
SELF LEARNING ENGINE
SLE (SELF LEARNING ENGINE)
Consumer

ASSESSMENT RULE ENGINE

Examination and initial evaluation engine

- Contains all the relevant laws and standards relevant to the case and DI

SCHEMA BASED REASONING ENGINE

- A complex case is transferred to the Scheme Based Reasoning Engine (SBR), which recognizes the relevant parameters to make the decision, examines possible results and their probabilities.
- The SBR transfers the subject – with a suggestion of alternatives for the decision – to the Case Based Reasoning Engine (CBR).

CBR
CASE BASED REASONING SYTEM

- Study similar cases from the past as a precedent , add probabilities where needed to the decision making process and return for final decision at the Assessment Rule Engine

(SLE)
SELF LEARNING ENGINE

- Addition of the new cases and determination of new rules for analysis

- **Stage A:** the consumer approaches the system and requests a service through the Internet (or scanned paperwork) and its content is 'read' by an 'engine' able to interpret texts;
- **Stage B:** the system checks the consumer's digital identity (DI) and registers it with all its components from the digital identities database;
- **Stage C:** the request is combined with the consumer's digital identity and is directed to the Assessment Rule Engine (ARE) that contains all relevant laws, standards and legal precedents; the system examines the request and—if it is a simple case in which all relevant parameters are known—the consumer receives an immediate decision;
- **Stage D:** a complex request without 'simple case' precedent in the database, may require elements of uncertainty, is directed to the Schema Based Reasoning Engine (SBR); the SBR identifies the relevant decision-making parameters, examines possible results and their probabilities and transfers the matter—with suggestions of alternatives for the decision—to the Case Based Reasoning Engine (CBR);
- **Stage E:** the CBR engine, which has access to a database with similar previous complex cases, automatically selects the final components of the decision (alternatives and probabilities) and transfers the matter to an Assessment Rule Engine for the final calculation and decision;
- **Stage F:** the ARE engine re-examines the decision, considering the laws, standards and precedents, and makes a decision that is presented to the consumer. In addition, the decision is transferred to the SLE (Self Learning Engine), a database of the decisions made, for continuous self-learning and updating probabilities; the whole process is completed at the speed of light.

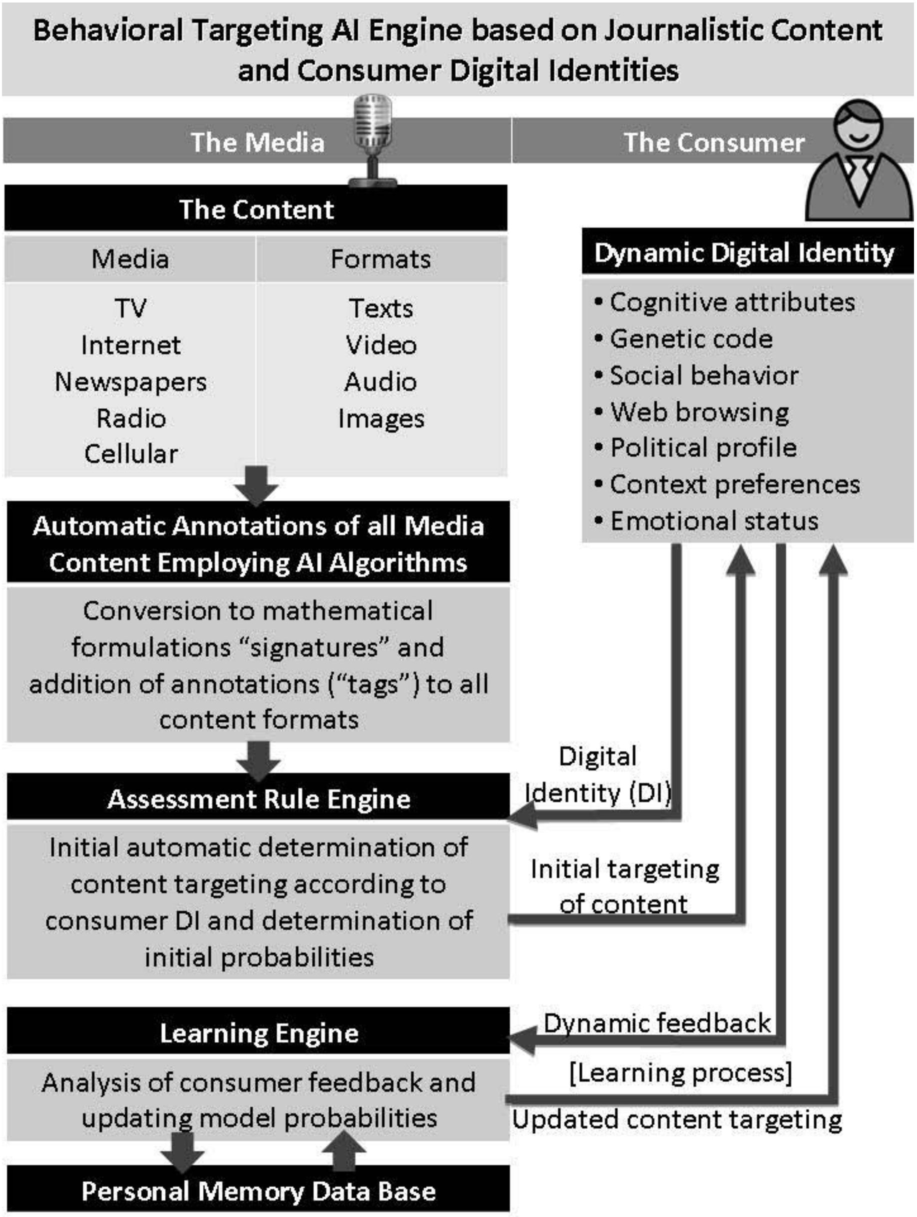
The consumers' digital identities are a vital component in this process and will directly affect the type of services and information they will receive in their lifetime.

Several countries and governmental authorities are adopting the automated decision-making processes shown in the flow chart. Australia is implementing these processes in several governmental bodies: its Ministry of Agriculture uses decision engines such as these to decide whether to allow an import, whether to supervise imported merchandise and which tests to perform; its Ministry of Finance uses artificial-intelligence engines to determine taxes, bonuses and fines; its Ministries of Health, Defense and Immigration are assisted by artificial intelligence for automated decision making to check cargoes at ports of entry.

In the United States the system is called 'Automated Targeting Spotter.' Its goal is to identify cargoes with high risk for terrorist activities; the system includes more than 300 'rules' written by teams associated with the field.

Will replacing human intelligence in decision making with a computerized model, as sophisticated as it is, bring better results, improve personal freedom and make society's ability more efficient to properly exploit its resources? These questions will be answered over time. Decision quality will be influenced by the efficiency of the model's structure used, the nature of the interface between the various engines during the decision process, the levels of their updating, the strength and accuracy of the basic assumptions made by the algorithm engineers and—most important—system 'noise' level.

6.3 Behavioral Targeting AI Engine Based on Journalistic Content and Consumer ID


Behavioral Targeting AI Engine based on Journalistic Content and Consumer Digital Identities
The Media
The Consumer
The Content
Media
Formats
TV
Internet
Newspapers
Radio
Cellular
Texts
Video
Audio
Images
Dynamic Digital Identity
• Cognitive attributes
• Genetic code
• Social behavior
• Web browsing
• Political profile
• Context preferences
• Emotional status
Automatic Annotations of all Media Content Employing AI Algorithms
Conversion to mathematical formulations "signatures" and addition of annotations ("tags") to all content formats
Digital Identity (DI)
Assessment Rule Engine
Initial automatic determination of content targeting according to consumer DI and determination of initial probabilities
Initial targeting of content
Learning Engine
Dynamic feedback
Analysis of consumer feedback and updating model probabilities
[Learning process]
Updated content targeting
Personal Memory Data Base

The behavioral targeting AI engine described above outlines the basic information-flow elements that will automatically analyze journalistic content in all platforms and transmit the relevant content and advertisements to consumers per their digital identities. The basic architecture of this model is similar to the AI engine used by governments as described earlier.

A brief description of the information flow in the diagram:

1. All journalistic content is analyzed by AI smart algorithms and receive automatic annotations (tags);
2. Consumers' digital identities and annotated content are fed to the Assessment Rule Engine for initial content determination; proper ads and content units are sent to consumers based on their profiles;
3. Consumers interact with the content and advertisements, this interactivity is being monitored constantly and consumer attention measured;
4. The Learning Engine analyzes consumer feedback and automatically adjusts the probabilities to better describe consumer behavior; new content units are sent to the consumers;
5. The Learning Engine transmits updated information to a Personal Memory data base where a consumer media profile is created and constantly updated;

Steps four and five continue indefinitely to allow the AI engine to accurately predict consumer content and product interests/choices in varying contexts—the "Learning Process" section.

As the model shows, it is a dynamic learning model constantly updated as it 'learns' the consumer profile and content preferences. Unknown factors are expressed by probabilities constantly updated in the 'learning' process. Journalistic content will be monitored constantly as consumers interact with it and make choices. The AI engine will also monitor context parameters and consumers' emotional state during interaction by analyzing their verbal or other reactions.

Some major concerns arise from the expected increase use of the automated decision-making processes described above;

Does a certain person's digital identity indeed describe this person correctly? Is it acceptable to allow or prevent a person's accessibility to content according to the digital identity that people cannot control? Do the implemented international standards re digital identities answer this important issue? Will it be allowed to include genetic aspects as part of the digital identity which may determine the fate of a person at birth? And an additional important question: will it be legal to include probability components in describing a person's digital identity (as can be expected to be constructed in a decision- making model)?

Filtering journalistic content vs. consumer profiles could lead to serious social inequality. Marcia Sterpanek coined the term 'weblining' to describe this phenomenon: "Call it weblining, an information age version of that nasty old practice of red lining, where lenders and other businesses mark whole neighborhood off limits. Cyber space doesn't have real geography but that's no impediment to weblining [...] weblining may permanently close doors to you or your business."[82]

The University of New York sociologist Marshall Blonsky adds to the meaning of 'weblining:' "If I am weblined and judged to be of minimum value, I will never have the product and services channeled to me or the economic opportunities that flow to others over the net."[83]

The digital identity is at the core of weblining. Though the emphasis of Stepanek and Blonsky is on economic aspects of commercial organizations, the described phenomenon is also true in spreading journalistic content based on profiling. The economic forces—advertisers and the journalism organizations—cannot be expected to show altruism and create mechanisms to protect our right to equal accessibility to content.

7 Digital Identities and the Practice of Journalism

From the point of view of journalism practice, the emergence of digital identities suggests that publishers and journalists will be able to simulate and measure what their news stories will do for audiences and the other stakeholders in their storytelling, while they are developing the story. They would be able to 'test run' stories before publication, much as advertisers now do with new product tests. This will introduce interesting opportunities and challenges for journalism.

Simple on-textual advertising need not threaten journalistic principles of separation between content production and selling audience attention to advertisers (who might have stakes in the journalistic stories). But as contextual advertising starts to understand the content, context and audience better, ads will be placed with surgical precision. Present pay-per-click business models will create an incentive for news publishers to focus on stories that will match ads. If that happens, it will threaten journalistic freedom. The classic 'separation of Church and State,' the metaphor used historically by publishers to distinguish between news stories and paid advertising, will blur.

For example, consider a situation where readers use their digital identities, combined with a series of filters, to select news stories they want to be brought to their attention. Let's say the quality of filters and digital identities is good enough to estimate both the chance that a story will catch the reader attention and the

[82] Sterpanek, Marcia. **Weblining**, businessweek on-line, April 3, 2000.

[83] ibid.

chance it will lead to action by the reader. Now consider a set of contextual advertisers (these can also be digital identities) that will pay for attention and interaction with readers. Consider a journalist with access to these digital identities and filters, as well as access to the contextual advertisers, when writing a story. The journalist can test the story on digital identities representing both audience and advertisers as the story is written. The journalist can adjust the writing to receive the 'best' results, a combination of what the journalists, the audience wants and the advertisers want.

Consider, finally, that the journalist's own digital identity will be included in the interaction, The journalist's digital identity is combined with a set of filters for selecting themes that the journalist wishes to cover, connected to readers' and advertisers' digital identities, and exposed to a 'news ticker' type flow of events, e.g. all the twitter feeds, the blogosphere and all the other news feeds on the Internet. It can be data flow from stock markets, sensors measuring weather or earthquakes etc. The journalist can then be tipped off about events that will produce suitable matching between his/her own interests, and the interests of the audience and advertisers.

Thus producing a successful story is equal to solving a dynamic equation involving the journalist, the audience and the business model, e.g. the advertiser. Producing a journalistic story while guided by the interaction between the digital identities and the filters can be seen as an iterative, heuristic solution of the equation, identifying overlapping interests and optimizing the combined actions into a result maximizing value for each interacting party. In each interaction, the real-life users behind the digital identities give feedbacks, reinforcing or modifying the behavior of the digital identities and filters, to improve the outcome in the next round.

7.1 Concern: Principles of Journalism @ Digital Identities

The interaction between digital identities, as discussed above, may improve the outcome for all parties involved. But it is a hazardous scenario. It needs to be discussed among the actors who care about journalism and its role in society.

Looking at existing journalism principles, at least the following can be strongly affected by the above scenario:

- **Journalism's first loyalty is to the citizens:** Journalists can be pressured to show loyalty to citizens' digital identities rather than to the citizens themselves. If each story is coupled directly to the business model, and if the business model builds on selling audience attention/interaction to advertisers, this can be a problem. It will be difficult to maintain a loyalty to the audience of citizens if the journalist will earn more money by adapting to the [digital identities of the] advertisers.

- **Its practitioners must maintain independence from those they cover:** It may be possible to involve behavioral models of those covered in the stories in the 'equation.' This will improve the possibilities for journalists to plan a series of stories, knowing how the outcome of one story will open for the next. It will give journalists a tool for projecting the effects the story will have on its stakeholders. Those covered in the story may also be advertisers, or have strong, shared interests with advertisers. This makes the wide web of co-dependencies much more visible to the journalist. In some cases this can help a journalist to be independent but in many other cases it will make it difficult to maintain independence.

- **Its practitioners must be allowed to practice their personal conscience:** If the business model and the system of digital identities and filters makes it possible to project how much profit a certain story can produce as the story is written, or if it will offer predictions of how the story will influence stakeholders in the journalism organization, probability increases that the journalists' personal consciences may conflict with business interests or other stakeholder interests. In short: 'if I write the story the way I want it, my publisher will know that I actively chose to earn less money'. Or: 'If I write the story the way I want, my publisher will know that I actively chose to increase the risk of us getting in conflict with NNN.'

These are only quick, simple examples of types of issues that need to be considered while developing systems of digital identities and filters for journalism.

A group of computer scientists, artificial intelligence researchers and roboticists met in Asilomar Conference Grounds on Monterey Bay in California to debate "whether there should be limits on research that might lead to the loss of human control over computer based systems that carry a growing share of society's workload…their concern is that further advances could create profound social disruptions and even have dangerous consequences…and force humans to learn to live with machines that increasingly copy human behaviors"[84].

The scientists were concerned about job loss or criminals accessing these tools. No reference was made to the possible devastating effects that the use of AI tools may have on journalistic content. The conference was organized by the Association for the Advancement of Artificial Intelligence(AAAI). Dr Horvitz of Microsoft, who organized the meeting, said "he believed computer scientists must respond to the notions of superintelligent machines and artificial intelligence run amok …the panel was looking for ways to guide research so that technology improved society rather than move it toward technological catastrophe"[85].

It is time to organize a similar conference with computer scientists, AI experts, academic researchers in the area of multimedia information retrieval, social communication experts and economists who specialize in media business models, to explore the potential effects of AI algorithms on the journalism profession and its role in a democratic society. Questions to be raised include:

[84] Markoff John, "Scientists Worry Machines May Outsmart Man" ,NYT.com july 26 09

[85] Ibid

1. Will people control or be controlled by their digital identities?
2. How will the definition of journalism be influenced by digital identities?
3. With the introduction of the Internet, journalism is no longer only broadcasting stories, but also interacting with readerships and facilitating public discussions. What is the role of journalism in the information society?
4. How will the existing principles of journalism be affected by the interaction between digital identities?
5. Which business models are enabled by digital identities? To what extent will journalists be attention workers, paid by brokering the readership attention to advertisers; to what extent will they be knowledge workers, paid by brokering knowledge?
6. What are suitable principles for journalism, in a situation where interaction with and between digital identities guides the production of journalism, the ways it generates value for people, and the ways it creates profits for the journalism industry?
7. What is the match between journalism and journalistic business models?
8. How will journalistic principles and matching business models be updated?
9. How are journalistic principles, and the process for updating will be implemented in an environment of digital identities?

Noam Lemelshtrich Latar is the Founding Dean of the Sammy Ofer School of Communications at IDC.Herzliya (the first private academic institution in Israel), and serves since 2009 as the Chairperson of the Israel Communications Association, which groups all media researchers in the Israeli Universities and Colleges. Dr Lemelshtrich Latar received a Ph.D. in communications from MIT in 1974 and an M.Sc. in engineering systems at Stanford in1971. He was among the founders of the Community Dialog Project at MIT, experimenting with interactive TV programs involving communities through electronic means. From 1975 to 2005. Dr. Lemelshtrich Latar pioneered the teaching and research of new media at the Hebrew and Tel Aviv Universities. From 1999 to 2005 he was involved in the Israeli high-tech industry as a venture-capital chairman, helping to establish several communications start ups in cognitive enhancement, data mining of consumer choices and home networking. In 2005 he joined IDC Herzliya Isreal as founding Dean of a new school of communications, emphazing new media. His current research interest is in digital identities and AI decision-making in 'E' government.

David Nordfors is co-founding Executive Director of the VINNOVA Stanford Research Center of Innovation Journalism at Stanford University. He coined ' Innovation Journalism' and 'Attention Work' and started the first innovation journalism initiatives, in Sweden and at Stanford. He is a member of the World Economic Forum Global Agenda Council on the Future of Journalism. Nordfors is adjunct professor at IDC Herzliya and visiting professor at the Monterrey Institute of Technology and Higher Education (Tech Monterrey). Dr. Nordfors has a Ph.D. in molecular quantum physics from the Uppsala University, and did his post-doctoral research in theoretical chemistry at the University of Heidelberg. He was the initial Director of Research Funding of the Knowledge Foundation in Sweden (KK-stiftelsen). He was the first Science Editor of Datateknik, a Swedish IT magazine, from where he initiated and headed the first hearing about the Internet to be held by the Swedish Parliament.