Society for Philosophy and Technology

Techné:

Research in Philosophy and Technology

Joseph C. Pitt, Editor-in-Chief
Pieter Vermaas, Editor
Peter-Paul Verbeek, Editor

Volume 11 Number 1 Fall 2007

Technè: Research in Philosophy and Technology

CONTENTS

JOSEPH C. PITT, PIETER VERMAAS, and PETER-PAUL VERBEEK, Editorial Statement 1

PHILIP BREY, Theorizing the Cultural Quality of New Media 2

PAUL THOMPSON, Theorizing Technological and Institutional Change: Alienability, Rivalry and Exclusion Cost 19

JOSEPH REAGLE, JR., Bug Tracking Systems as Public Spheres 32

PETER KREBS, Virtual Models and Simulations: A Different Kind of Science? 42

JONAS CLAUSEN and JOHN CANTWELL, Reasoning With Safety Factor Rules 55

BERNADETTE BENSUADE-VINCENT and XAVIER GUCHET, Nanomachine: One Word For Three Different Paradigms 71

Techné: Research in Philosophy & Technology proudly announces a new editorial staff, consisting of Joseph C. Pitt of Virginia Tech as Editor-in-Chief, Pieter Vermaas of Delft University of Technology and Peter-Paul Verbeek of University of Twente as Editors, Tom Staley of Virginia Tech as Book Review Editor and Ashley Shew of Virginia Tech as Managing Editor. The Editors welcome submissions of all styles and approaches in philosophy of technology and look forward to expanding the scope of the journal while increasing standards of the work published.

Techné welcomes all philosophical perspectives and styles. The editorial stance of *Techné* is ideologically neutral. There is, however, a unifying theme: a focus on technology, particular technologies, modern or traditional, worldwide, or on social and ethical problems associated with particular technologies. *Techné* aims at being the platform for presenting novel developments and results in academic research on this theme. We therefore seek rigorous, seminal, interesting, creative work and eagerly solicit work from those in fields outside philosophy as long as they offer philosophical perspective.

Submissions will be blind refereed by at least two readers. It is our expectation that authors will be provided with critiques, where, in the judgment of the editors they are deemed helpful. We also seek to have a turn-around time of two months, although this is subject to the cooperation of our referees.

We construe philosophy of technology broadly, and we are dedicated to fostering the highest standards in what is becoming a diverse field of study. We hope you will consider submitting your work to *Techné*. Submissions and questions can be directed by email to technejournal@gmail.com. If you would like to read past issues of *Techné*, please visit our Ejournal page, http://scholar.lib.vt.edu/ejournals/SPT/.

Theorizing the Cultural Quality of New Media

Philip Brey
Department of Philosophy
University of Twente

1. Introduction: Normative and Critical Studies of New Media

The past thirty years have witnessed the emergence of new media: interactive, computer-based devices like multimedia PCs, digital (mobile) telephones, the Internet, hand-held computers and game computers. All of these are made possible through new advances in information technology. These devices are now regularly used at work or at home by a majority of people, and their influence has extended deeply to all sectors of society, including work, leisure, education, health care, government and the arts. New media have become new mass media, contrasting with "old" electronic and print media, like the radio, television, telephone and newspaper. It is widely recognized that the social, cultural and political implications of new media are significant, and it has even been argued by many that their rise has enabled the emergence of a new, postindustrial model of society, the information society, with its own principles of social and economic organization and cultural practices (Castells, 1996). The social, cultural and political implications of new media have now become a major topic in academic research, in both the social sciences, humanities and arts. In recent years, an interdisciplinary field of new media studies has even emerged (Lievrouw & Livingstone, 2002; Lister et al., 2003; Wardrip-Fruin & Montfort, 2003).

Whereas most research on new media is descriptive and empirical, part of it is normative and evaluative: it proposes normative criteria for the evaluation of social and cultural implications of new media use and evaluates such implications so as to assess their value, desirability, quality or worth. A survey of such normative analyses shows three major traditions: ethical analysis, normative political analysis, and aesthetic analysis.[1] Ethical analysis of information technology has been the province of the field of computer ethics (Johnson, 2000; Tavani 2003). Computer ethics is a field that emerged in the 1980s out of worries stemming from unethical uses of computers, for instance in computer crime, privacy violations, free speech and censorship, and property rights, and is mainly concerned with the analysis right and wrong conduct in the use of computer technology and the formulation and justification of policies for its ethical use. Normative political analysis, which sometimes overlaps with ethical analysis, develops conceptions and arguments concerning the role of the state in the regulation of computer technology, the role of the law in such regulation, and the distribution of powers and responsibilities between citizens, corporations and the state in the use of such technology (Hill and Hughes, 1998; Saco, 2002; Mossberger et al., 2003; Fountain, 2001). It discusses issues like cyberdemocracy, distributive justice and the digital divide, liberal vs. conservative policies for regulating free speech on the internet, the protection of property rights, national security, and the common good in cyberspace. Aesthetic analysis finally, discusses the aesthetic and literary properties of new media creations and evaluates the impact of new media on our conception of art and literature (Gumbrecht and Marrinan, 2003; Wardrip-Fruin and Harrigon, 2004; Dyson and Homolka, 1996).

[1] Epistemological analyses could constitute a fourth class, but very few epistemological studies of new media exist.

Critiques that Do Not Fit the Mold

There is a growing number of studies that critically examines the social and cultural impacts of new media that cannot be seen to fall in any of these three established categories, the ethical, the political, or the aesthetic, even though they voice normative and evaluative criticism:

- In *Holding on to Reality,* Albert Borgmann develops a critique of cyberspace (Borgmann, 1999). Borgmann argues that cyberspace presents an illusory escape into another reality. He claims that it tends to trivialize and glamorize facets of reality that appear to one detached from their context and setting, and that it blurs the distinction between fact and fiction.

- In discussing the implications of cyberspace for identity, Sherry Turkle has argued that the multiple virtual personalities that people may adopt on the net may promote the emergence of a nonunitary, multiple self. This, she says, can be evaluated negatively if one adopts the ideal of a unitary, autonomous, modernist self, but which is evaluated positively by her because being able to emphasize different aspects of oneself in different identities can be liberatory and can help us better acknowledge diversity (Turkle, 1995).

- Hubert Dreyfus has critiqued computer-mediated education (Dreyfus, 1999). He argues that education centrally involves the transmission of skills and a process by which educators foster commitments in their students and stimulate them to develop strong identities. He then argues that such skills, commitments and identities cannot adequately be transferred in distance education since they require bodily presence and localized interactions between students and teachers. This requires a relation of apprenticeship, which according to Dreyfus cannot be attained on-line.

- Paul Virilio has argued that electronic media, developed and used in a capitalist consumer society, combine with other technologies in speeding up the process of production and consumption so as to create a culture of speed (Virilio, 1994). The immediate availability of information and the continuous production and consumption of new information ultimately lead, according to Virilio, to a feeling of confinement or incarceration in the world. Virilio also holds that the culture of speed threatens writing and the author, because the speed with which information is produced and consumed only allows for shallowness.

- Langdon Winner has argued against a conception of virtual communities as real communities, arguing that most of them do not include the obligations, responsibilities and constraints found in ordinary communities, while they may well end up undermining real communities, which makes all of us lose (Winner, 1997).

- Ben-Ze'ev, finally, has argued that cyberspace has revolutionized the role of imagination in personal relationships by coupling imagination with real interactivity, allowing us to have meaningful online relationships in which we can both express ourselves in more direct ways than we would otherwise and live out fantasies, which he evaluates mostly in a positive way (Ben-Ze'ev, 2004).

None of these critiques, I want to claim, fall clearly within the traditional categories of ethical,

political or aesthetic analysis. So what kind of critique are they? The most obvious answer seems to be that they are *cultural critiques*, since they seem to have as their object cultural practices, symbols, meanings and configurations; that is, they critique culture. While, I will admit, one could describe them as cultural critiques, such a description is not sufficient in distinguishing them from other types of normative critique. The problem is that a "cultural critique" may simultaneously also be political, ethical, or aesthetic. There is a long tradition, dating back to Marx and the Frankfurt School, of cultural critique as political critique, and with the exception of occasional aesthetic critiques of the natural world, all aesthetic critiques are also cultural because they are directed at products of culture. Critiques of culture can also be ethical critiques. For instance, it can be and has been argued that a corporate culture that promotes greed leads to morally impermissible behavior and is therefore wrong.[2] So if the above critiques do not form a distinct class of cultural critiques, how can the instead be categorized? Or do they not form a distinct class at all? In the next section, I will argue that they do form a special class, and that recognition that this is the case will help this kind of criticism gain a higher profile and create more coherence and dialogue in the area of research in which these critiques can be located.

2. Theories of the Good and their Relation to Culture

The main question raised in the previous section is whether the examples that were given constitute a particular type of normative critique. Let me try to answer the question by asking how certain normative questions and issues gain coherence and become recognized as separate fields. This occurs, I claim, when they are centered around a particular normative ideal that is valued widely. Ethical analysis is concerned with the Right: it is based on a drive to understand what kinds of actions are right and therefore obligatory, and which ones are wrong and therefore impermissible. Normative political analysis is concerned with the Just: it is based on a drive to understand how the state ought to operate in relation to its citizens and how it should distribute powers and goods. Aesthetic analysis is concerned with the Beautiful, where "beautiful" is our most general term to express that something is pleasing or moving to observe.[3]

Could it be argued that the cultural analyses of the sort discussed above are all governed by a similar sort of ideal? I believe that this is the case, not because they are governed by a specific ideal, but because they are governed by our most general ideal, which is the Good. "Good" is our most general term of positive evaluation, and in philosophy a *theory of the good* specifies what sorts of things in life are good and therefore worth striving for (Ross, 1930; Larmore, 1996). What these examples therefore have in common is two things: (a) they critique culture; and (b) they do so in light of an ideal of the Good. That is, they employ some conception of what would be good or bad for individuals or for society and they criticize cultural developments in light of this conception.

Theories of the good have a long history in philosophy, beginning with Plato's view that the good is the principle of reality and Aristotle's assertion that the goodness of things is determined by the question of how well they live up to their final cause or function. Aristotle also famously claimed that the good for human beings is found in the cultivation of human virtues, being human

[2] It will also not do to argue that some of these critiques are metaphysical or epistemological in nature. Borgmann and Baudrillard may be claimed to be tackling metaphysical issues, but their ultimate aim is not to investigate reality; it is to critique worrisome changes in our relation of and conception of reality that they believe have bad cultural effects. Likewise, Dreyfus discusses epistemological issues in his critique of learning, but is more broadly concerned with the transfer of skills and academic values.

[3] Epistemological critiques, one may add, are concerned with the True and with justifying that our beliefs are true.

capacities that are part of their final cause, and particularly in the cultivation of rationality. This would result in the highest good for human beings, which he called *eudaimonia*, or personal flourishing. In the modern era, theories of the good have often been developed in the context of consequentialist ethical theory, particularly in utilitarianism. These have resulted in various sorts of hedonist and preference-satisfactionist theories of the good. The recent revival of Aristotelian virtue ethics has also resulted in new varieties of virtue-based (or perfectionist) conceptions of goodness and the good life (Nussbaum, 1986; Hurka, 1993).

Theories of the good have a somewhat ambiguous status in philosophy. They are usually considered to be part of ethics, specifically normative ethics, which has traditionally been defined as consisting of a theory of the good and a theory of the right. Normative ethics is then held to have two tasks: to develop a theory of the good, which specifies what is good and therefore worth striving for, and to develop a theory of the right, which directs itself at human action and aims to determine which actions are right or wrong. In popular conceptions of normative ethics, however, normative ethics is primarily if not exclusively concerned with the *rightness or wrongness of actions*, and theories of the good are not conceived of as having a separate status in ethics. At best, such theories are then conceived of as prerequisites to the development of a theory of the right. This is a particular necessity for consequentialist theories of ethics, such as utilitarianism, which as they evaluate the morality of actions by the goodness or badness of their consequences, and therefore must be grounded in a theory of the good. Deontological theories, in contrast, hold the right to be prior to the good and therefore do not require a theory of the good. Virtue ethics, as a third major type, grounds both a theory of the good (*eudaimonia*) and of the right (virtuous action) in the notion of a virtuous character. A virtuous person is twice lucky: he or she is compelled to behave morally and he or she has well-being because of one's balanced character.

Theories of the good are sometimes also placed under the heading of the theory of value, or axiology, which is a branch of philosophy concerned with a general analysis of value or quality. Ethics and aesthetics are sometimes even classified as the two major branches of axiology. Goodness is of course itself a value, like beauty, rightness and justice, and it can even be claimed to be our highest term for evaluation. It is fair to say that theories of the good are located at the intersection of ethics and theory of value. However, ethics is often defined narrowly as the study of morality, or of right and wrong action, which would exclude an independent consideration of (nonmoral) goodness. It is therefore perhaps better to categorize studies of the good as a separate branch of axiology or theory of value (Carson, 2000; Rescher, 2004).

What critical discussions of new media show, more than anything, is that the transformative effect of new media on human culture is so profound that general questions about the good are being raised that cannot be answered in terms of the more narrow categories of ethics, politics, or aesthetics (Brey, 1998). What I am therefore proposing is the development of an applied area of research where theories of the good are applied and developed in relation to new media and new media culture. Framing existing cultural critiques of new media, such as the ones discussed in the previous section, in this way will make it possible to relate them to the general and explicit accounts of the good that have been developed in philosophy over the course of several thousand years. Current cultural critiques of new media often leave their conceptions of the good implicit, and rely on an intuitive recognition of the validity of their critiques in the reader. Partially because of this, studies in this area lack unity and a common vocabulary, making reasoned discussion of and comparison between them difficult. It would therefore be better if the conceptions of the good used in these critiques could be made explicit and could be discussed in

relation to existing accounts of the good in philosophy. This could both lead to a better understanding of such critiques and facilitate comparison of and dialogue between them.

In the remainder of this section, I will review major theories of the good that have been developed in philosophy, after which I will discuss how such theories may be useful in constructing a theory of the goodness of culture. In section 3, I will then go on to analyze how theories of the good may be applied to the analysis of technology in general, and more specifically to new media and new media culture.

The Good as the Human Good: Theories of Well-Being and the Good Life

The question "what is good?" is often understood under an implicit assumption that human beings are the measure for goodness, and is then interpreted to mean "what is good for human beings," which is then translated as "what is the good life" or "what is well-being"? Most theories of the good, therefore, are actually theories of the good life or well-being: they assume that only good lives have intrinsic worth, and the things we call good are things that contribute to a good life, which is a life in which individuals have well-being (welfare, quality of life). Well-being is a kind of value which is sometimes called "prudential value," a type of value that exists alongside aesthetic and moral value, amongst others, and which has the characteristic property of being *good for* someone. It is generally recognized in philosophy that there are three major types of theories of the good life: hedonist, desire-fulfillment and objective theories (Parfit, 1986).

Hedonist theories hold that only pleasure is intrinsically good, and pain is the only intrinsic bad. Several varieties of hedonism exist, including *quantitative hedonism* (or simple hedonism), which holds that the value of pleasure is only determined by its duration and intensity, and *qualitative hedonism*, which holds that some pleasures (for instance those related to contemplation and intelligence) are more valuable or pleasurable than others. One prominent objection to hedonism has been proposed by Robert Nozick, who hypothesis an "experience machine" that simulates a nonexistent world in which one has all experiences of whatever kind one finds most enjoyable (Nozick, 1974). Many agree that it would be undesirable to plug in to such a machine, since one's experiences are not based on actual events but on simulations, and therefore less valuable.

Desire-fulfillment theories, also called *preference-satisfaction theories*, hold that well-being lies in the fulfillment of one's desires. They are favored by some over hedonism because they are capable of avoiding the "experience machine" dilemma: if one desires to be loved by friends, and an "experience machine" simulates loving friends, then one's desire is not fulfilled, and this experience is therefore less valuable than one that really fulfills one's desire. A major impetus for the development of desire-fulfillment theories instead of hedonist theories has been economists looking for a more objective measure for welfare. Happiness and pain are, after all, in the head, and cannot easily be measured. Statements about one's preferences, and the rankings one assigns to them, can be more objectively determined. A distinction can be made between *simple desire-fulfillment theories*, which merely hold that the best life is the life in which all one's desires are fulfilled, and *informed desire-fulfillment theories*, which holds that the best life one could lead is the life in which all desires are fulfilled that one would have if one were fully informed of one's situation.

Objective theories, which have also been called *objective list* theories, hold that well-being is the result of a number of objective conditions of persons rather than the subjective experience of pleasure or the fulfillment of their subjective desires. They propose that some things contribute

to our well-being even if they do not give us pleasure or correspond to our desires. Conditions that have been proposed as part of such a list of conditions include knowledge, friendship, the development of one's abilities, having children and being a good parent, the awareness of true beauty, and moral goodness. *Perfectionism* is an influential kind of objective theory that proposes that what makes things constituents of well-being is their perfecting human nature. On this conception, humans are held to have a *telos* or end that can be attained if the right conditions are met. When they are met, the person has attained a state of well-being. One famous perfectionist theory is Aristotle's theory of *eudaimonia*, as mentioned previously.

Other Conceptions of the Good

Theories of the good that hold that the only intrinsic good is individual well-being may be called *individualist*. Although they have not drawn much attention in philosophy, other conceptions of the good are possible, and instead can be located in various value systems, that (also) hold things to be intrinsically good that do not contribute to human well-being. Two general kinds of theories may be distinguished, which I will call collectivist and transcendent. *Collectivist theories* hold that the greatest good is not the good of individual human beings but the good of the larger collective, such as a tribe, a community, or society at large. It is doubtful however, that existing ideologies that emphasize the common good or the good of society, like communitarianism, socialism and communism, truly hold that only the good of the larger collective is intrinsically valuable. These ideologies usually make the additional claim that the good of society is only a shorthand for the good of the individual members of society. When this additional claim is made, these ideologies turn out to be individualist after all. They only disagree with more liberal ideologies about the best way to realize the good for individuals, arguing that this is to be attained through promotion of the common good.

Transcendent theories, finally, hold that humans, whether as individuals or as collectives, are not the measure of goodness. Such theories point to one or more transcendent state-of-affairs or qualities that are held to constitute the highest good. Alternatively, they may hold that certain things are intrinsically valuable in addition to, and independently of, the human good. Transcendent goods that have been proposed include the glory of God or obedience to God's law (in Christianity and Judaism); the natural order of things (Taoism); the realization of its *telos* by all of life; ecosystemic integrity, or the well-being of Gaia (mother earth, conceived of as a living being); truth, knowledge or information (e.g., Floridi, 2002); artistic or natural beauty (radical varieties of aestheticism). Some mystics also hold that the universe has a purpose or value according to the will of a creator, but which lies beyond human understanding.

Conceptions of the Good and Cultural Quality

As a next step, I will now consider how theories of the good apply to culture. I will explore this by first defining culture and then analyzing how aspects of culture can be evaluated based on particular conceptions of the good. Encyclopedia articles on culture usually begin with a discussion of various definitions of culture that have been proposed over time. "Culture" is indeed a vague and ambiguous concept. What the various definitions of culture have in common is a recognition that culture is human-made, that it is learned or acquired, that it is shared by the members of a society, and that it is transmitted by nongenetic means from generation to generation (notably, by learning). There is also considerable agreement that culture is an adaptation mechanism that enables societies to better adapt to the environment and to maintain social order and stability.

In addition, there is a rather broad recognition that culture is made up of at least three types of entities: symbols, behaviors and artifacts. Symbols are arbitrary signs used to convey meaning, and whose meaning is determined by social convention and learning. They include human language, symbolic gestures, symbolic images and markings, and all kinds of nonlinguistic signs in artifacts, like traffic signs, flags, or crucifixes. Cultural behaviors or practices or customs are socially learned actions or scripted patterns of action that may or may not involve specific settings and artifacts, and may be individual or collective (refs.) Most behaviors, ranging from the way one holds a cigarette in one's mouth to eating with a fork and a knife to courtship and marriage, are strongly conditioned by social learning. Artifacts, finally, are products of material culture: they are human-made material goods like clothing, furniture, tools, jewelry, artworks, and dwellings. Symbols could also be conceived of as artifacts, because they have in common with material artifacts that they are human creations that serve a purpose, and in a still broader sense, everything about culture can be considered an artifact, since it is a product of human making.

Other entities that are often held to be components of culture are beliefs, values, norms and institutions. Cultural beliefs are socially transmitted beliefs in a culture that may range from mundane beliefs about the poisonousness of certain berries to deeply held religious and metaphysical beliefs about the universe and one's place in it. Cultural values are shared, socially transmitted values that are ideals about what is important in life. They may, like beliefs, range from the mundane to the religious and metaphysical. Values can specify things that are valued (desirable behaviors, attitudes or conditions) or abstract ideals. Examples of cultural values are humbleness, rationality, honor, spirituality, efficiency, punctuality, individuality, happiness, peace, tradition, family closeness and professionalism. Norms, which tend to be related to a culture's values, consist of expectations of how people will behave in different situations. Norms may be formal or informal, and cultures have different methods, called sanctions, of enforcing their norms. Institutions, finally, are more or less permanent mechanisms of social structure that maintain social order by imposing and enforcing norms and corresponding cultural behaviors. Examples of institutions are the family, the state, law, religion, economic systems and the military. Institutions can be understood as nothing more than interrelated sets of norms and practices, and the mechanisms (artifacts, buildings, people) that are used to enforce them.

Taking these six elements of culture into account, we may now define culture as the system of shared symbols, behaviors, beliefs, values, norms, artifacts and institutions that the members of a society use to cope with their world and with one another, and that are transmitted from generation to generation through learning. This is a broad, anthropological definition of culture that is considerably broader and more profound than some more popular definitions of culture as more narrowly describe it as consisting of the arts and literature, or of the tastes in art, manners and lifestyle favored by a social group. It will be this broad, anthropological notion of culture that will be used in the remainder of this paper.

We may define a *theory of cultural quality*, or of the culturally good, as a theory of the good of culture in relation to some conception of the good. Such a theory would state the role or function or culture relative to the good, and the conditions that have to be satisfied by a culture for it to contribute well to some intrinsic good. Culture is a human-made artifact, or more precisely, a configuration of human-made artifacts, that may function either well or poorly. On most conceptions of the good, culture is an instrumental good: its function is to contribute to some higher good extrinsic to it, such as the satisfaction of individual desires or the good of society.

On an objective list or transcendent conception of culture, some aspects or products of culture, like knowledge or art, may also have intrinsic value.

Constructing a theory of cultural quality is difficult in the absence of an anthropological understanding of the functioning of culture and its specific cultural forms and artifacts in human societies. A theory of cultural quality has a different aim than an anthropological theory of culture that seeks to understand the proper function of culture in relation to human societies. Anthropological theories aim to give an objective account of the functioning of culture that is not guided by some normative conception of the good. Their aims are descriptive and explanatory: to understand why certain cultural forms have evolved and to understand their contribution to the functioning of society. It would be a naturalistic fallacy to translate such an anthropological conception of function into a normative conception of cultural quality: this would be deriving an "ought" from an "is". On a sociobiological conception of culture, for example, the function of culture is merely to help the human species adapt to its environment and propagate itself more successfully. It would obviously be wrong to say that because culture has historically had this function, we therefore ought to hold that the highest good of culture is its contribution to environmental adaptation and reproduction of the species.

Nevertheless, an anthropological understanding of the role of culture in human society is indispensable in constructing an account of cultural quality, because it gives one insight into what culture is and how it actually functions in human societies. Specifically, it may give insight into functions of culture that are not immediately obvious. The members of a society are normally only aware of what have been called the *manifest* or *conscious functions* of their culture: the functions that have been consciously and publicly assigned to it. Anthropologists, however, have tended to focus on *latent functions* of culture, which are functions about which the members of the community are not aware and which tend to benefit not individuals but the community as a whole (Merton, 1957). To evaluate the quality of cultural practices and meanings, an awareness of such latent functions is obviously needed.

Early functionalist accounts in anthropology held that culture was a collective means to satisfy individual (biological) needs. Bronislaw Malinowski, founder of the functionalist tradition in anthropology, held that the function of culture was to fulfill the needs of members of the culture (Malinowski, 1944). He held that humans have four basic biological needs that are common to all and that directly relate to survival, being the need for nutrition, safety, shelter and reproduction, and a larger number of derived needs that are culturally mediated or constructed, such as the need for psychological belonging to group, magic, religion and descent. Malinowski held that the satisfaction of these derived needs was ultimately to the benefit of the more basic needs. For example, he held that the extensive use of magic by Trobriand Islanders functioned to reduce their tensions and anxieties resulting form the uncertainties of life, which indirectly benefited their pursuit of their more basic biological needs. A functionalist account such as Malinowski's would fit well with an individualist account of the good, as it would imply that human culture already functions to promote human welfare, or some conception of it, so that the gap between the way culture functions and the way it ought to function may not be very large. However, functionalist accounts have largely been discredited as too much focused on the individual and insufficiently cognizant of sociocultural forces that transcend the individual.

Functionalism in anthropology has been succeeded by *structural-functionalism*, which was originated by A. R. Radcliffe-Brown. Radcliffe-Brown followed Emile Durkheim in proposing that a society is an integrated, organic system of interrelated parts that make a functional

contribution to the whole, and that culture exists for the benefit of communities rather than for individuals. Culture is held to have a social function, which is, in general terms, that it contributes to social order and equilibrium. Culture, in other words, functions to fulfill the "needs" of a social system, and not the needs of individuals. Of course, it is not denied by structural-functionalists that the contribution of culture to a well-ordered society could not also indirectly benefit individuals. Often, such benefits will materialize, but there is no necessity in this. A structural-functionalist account of culture would go well with a collectivist conception of the good, and would remind individualist theories that culture has a social function that, when neglected, could lead to social dysfunction and social instability, with possibly detrimental effects on individual well-being.

Much is still to be learned about the functions of culture in human society, including it various composite elements, like religion, language and art. Theories on these matters still often contradict each other, and are underdetermined by empirical evidence. However, it will be clear that an informed theory of cultural quality is dependent on anthropological studies on the functions of culture, and cannot neglect anthropology's best theories on this matter. Let me emphasize again that anthropological theories do not in themselves prescribe any conception of the good. A particular concept of the good, however, requires an anthropological theory of culture for a successful translation of this concept of the good to a theory of cultural quality. Such an anthropological theory can point to latent social functions of cultural forms and practices that have implications for one's normative conception of culture, as they give insight into the instrumental roles that culture does, can and must play.

3. The Good and New Media Culture

It may now start to become evident by now why new media are so much more transformative of culture than other modern technologies. New media, being media for information and communication, have become major carriers of cultural symbols; together with the "old" media, they have become culture's circulatory system. Even more so, in line with Marshall McLuhan's dictum that "the medium is the message" (McLuhan, 1964), it is undeniably so that new media are by no means neutral transmission media. They include new for techniques storing, representing, categorizing, transforming and communicating signs that have put their mark both the shape and the interpretation of cultural signs. Also, because they are interactive and capable of representing multimedial content and graphical environments, new media are used for much more than just communication and information transmission. They have become a medium for new individual and social practices and media in and through which institutions are realized (Mitchell, 1995; Brey, 2003). The technological infrastructure of new media even enforces norms though the structure of its hardware and software.

Having applied theories of the good to culture the previous section, I will now undertake a further application to technology, and then on to new media and new media culture.

Thick Conceptions of the Good and Comprehensive Doctrines

The theories of the good discussed in the previous section contain rather abstract proposals of particular notions of the good that are the product of the labor of philosophers. They are not, as such, actual conceptions of the good that are held and acted on by people in their everyday lives. Such conceptions of the good have been called *"thick" conceptions of the good*, which are

detailed systems of value that define what one finds valuable, including at least those things one finds intrinsically valuable, and possibly also one's conceptions of instrumental value, orderings between values, and one's attachments and loyalties to other humans, organizations and associations. Thick conception of the good are often part of more comprehensive ideologies or value systems that have been formed over time, possibly over centuries, that have attained some degree of institutionalization and that are shared by a larger group of people. I will call such ideologies, after Rawls (1993), *comprehensive doctrines.* (Another name may be worldview or ideology.) Comprehensive doctrines are systems of value, be they religious, moral or ideological, that contain a thick conception of the good which is often accompanied by norms for conduct and a system of (metaphysical) beliefs. People always have some thick conception of the good, and this conception may or may not be part of a comprehensive doctrine of which they are a follower. It is safe to say that people cannot develop a conception of the good out of the blue, and that their conceptions of the good, even if uniquely their own, are always indebted to one or more comprehensive doctrines to which they have been exposed.

Examples of comprehensive doctrines are world religions like Christianity and Islam and their different strands, and secular humanism. Religious systems often include a transcendent conception of the good (e.g. the glory of God, or obedience to God's law), but usually hold as well that humans have intrinsic value (for instance, because they are made in the image of God) and that their well-being is therefore important. Secular humanists do not recognize a God, and hold that the only good is the human good, implying that their highest good is individual well-being. The rise of a consumer society has lead scholars to characterize contemporary culture as a consumer culture, which carries its own set of values about what is important in life (Slater, 1997; Featherstone, 1991). Consumerism can be defined as an ideology that holds that physical well-being and the collection and consumption of material goods is the greatest good and highest value in life. In a secularized consumer society, it can be argued, advertisers have replaced the minister in advocating a particular conception of the good, or they are competing with him and winning. Consumerism can therefore be considered a new comprehensive doctrine being promoted by the modern market. Consumerism has been criticized because of its hedonism, individualism and self-interestedness, and its definition of the good life in terms of material goods, which critics have claimed should be considered instrumental goods rather than ends. Based in part on extensive empirical research, it has been argued that in the contemporary West, a new, postmaterialist doctrine is emerging in which people place greater values on ideas than on physical pleasure and material goods (Inglehart, 1990, 1997). Postmaterialists emphasize nonmaterial and nonhedonistic values like personal growth, quality leisure time, contemplation, meaningful relationships, care for the environment, social equality, and spirituality. The New Age movement can be seen as a manifestation of this, as well as the more recent voluntary simplicity movement, which embraces a lifestyle of lower consumption, less paid work, greater sustainability, less reliance on media technologies, and more self-reliance, which is argued to enhance the quality of life (Etzioni, 1998; Shaw and Newton, 2002).[4]

Political ideologies, like liberalism and socialism, are usually not comprehensive doctrines,

[4] It has been claimed that in a liberal capitalist consumer society, most people are no longer captivated by major comprehensive doctrines or ideologies, except for the general sort of consumerist attitude that comes with the culture. They are bestowed with considerable freedom to develop their own "rational life plans," as Rawls has called them, which has lead to the importance of developing one's own "lifestyle" in which personal values and beliefs become the basis for a way of life. Such lifestyles tend to become group phenomena that undergo a degree of institutionalization, in part because they often involve a consumptive element supported by commercial industries. One can often identify an ideological basis in them that can function as a limited kind of comprehensive doctrine. E.g., hippies, goths, bohemians, punks, yuppies, ravers, gamers, hackers (cf. Chaney, 1996).

because their aim is to specify the role of the state in realizing and distributing goods, and they often do so without advocating a particular thick conception of the good. However, sometimes they do presuppose a conception of the good, or at least a partial conception. Communitarianism, a political ideology that holds that the state should preserve communities and should often prioritize the interests of communities over those of individuals, presupposes a limited concept of the good according to which individual well-being is dependent on the well-being of communities. Communitarians have criticized the atomistic conception of the individual in libertarianism and liberalism, which seem to hold that well-being is an individual pursuit that can be defined without reference to one's membership in a community. Liberalism, in addition, famously employs a "thin theory of the good" according to which the principal task of government is to create the political and economic conditions under which individuals are freely able to pursue their own conception of the good (Rawls, 1971).

Conservatism, finally, can be understood as an ideology that strives to preserve existing social order and the institutions that sustain it. When these institutions embody a particular conception of the good, which is often the case, conservatism may take on the form of a comprehensive doctrine that seeks to uphold a particular conception of the good. However, different conservatisms may correspond to quite different concepts of the good. In Iran or China, traditional institutions embody ideals of the good that are quite different than those in the United States, so conservatism in these countries also means something different.

Studying Technology, Culture and the Good

Thick conceptions of the good, whether held individually or held collectively as part of comprehensive doctrines, find their first and foremost realization in the thoughts and behaviors of the people that hold them. However, they may also institutionalize and become embedded in a society's social structure and culture, including a society's customs, enforced norms, symbol systems and artifacts. Thus, a skyscraper is an artifact that is expressive of a particular value system, both in its symbolic meaning as icon of modernity, rationality, and transcendence, and in its compatibility with and support of particular practices and customs of modernity that are themselves in turn related to a particular conception of the good. The social and cultural shaping of human artifacts and technological systems is a central assumption in contemporary science and technology studies (STS), which took a constructivist turn in the mid-1980s and has since then been preoccupied with studying the social construction of modern technologies (Bijker, Pinch and Hughes, 1987; McKenzie and Wajcman, 1999). Yet, as is recognized in these studies, technology also has a role in shaping society and culture, although this role is always mediated by human action, and technology may have unintended consequences that are not compatible with the conceptions of the good and the intentions of those responsible for developing and using the technology. Technology hence embodies the values of a culture but may also affect culture in unintended ways and in divergence from these values.

Many studies in STS analyze political, cultural and aesthetic values embedded in technology and in the practices and meanings that have co-evolved with these technologies (McKenzie and Wajcman, 1999; Lievrouw and Livingstone, 2002; Misa, Brey and Feenberg, 2003), but so far these studies have not attempted to incorporate concepts and methods from the philosophical study of value and the good. In computer ethics, some philosophers inspired by STS have attempted to use concepts of ethics to develop approaches to the study of information technology that analyze the embedded moral values and norms in these technologies. This has resulted in

"values in design" approaches (Johnson, 1997; Nissenbaum, 1998) and "disclosive computer ethics" (Brey, 2000, 2001). These approaches have tended to focus on moral and political values and norms embedded in technologies that are analyzed in the context of theories of the right and normative political theory (e.g., values and norms relating to liberty, privacy, responsibility, democracy and justice) but have sometimes also considered values in the context of theories of the good (e.g., trust, community, or privacy and liberty understood as a components of a good life). What I propose here is to extend these approaches to focus more specifically and extensively on the good and the good life, rather than the ethics of obligation. Such studies will have to consider not just embedded values in technology, but also the embedded values in cultural practices, norms, symbols and institutions that co-evolve with these technologies and that may come to define them.

We may define an *axiology of technology,* or a *theory of values in technology*, as a general study of values embedded in technology, with an emphasis on those values that define notions of the good. An *axiology of technological culture* is an analysis of the practices, symbols, and other cultural forms that have co-evolved with specific technologies. An *axiology of new media* is a study of embedded values in new media technologies, and an *axiology of new media culture* (or *cyberculture*) is a study of values in the digital culture that has co-evolved with the rise of new media. A methodological assumption that I am making here, controversially, is that it is possible to perform an axiology of technology independently of, and prior to, an axiology of its co-constructed culture, that is, independently of its cultural embedding and use. Radical constructivists have argued that technology is social throughout, and that it makes no sense to speak of embedded values or inherent consequences in technology (cf. Brey, 1997). However, I propose to retain, as a working hypothesis, a very limited conception of autonomous technology according to which we can sometimes usefully refer to technologies as embodying values, meaning that technologies sometimes have normative consequences that do not co-vary greatly with their embeddedness in different social and cultural settings (Winner, 1980; Brey, 2005; Sclove, 1995). For example, I would want to hold that a web server that places and reads cookies on your computer is less protective of privacy than a web server that does not use cookies, or that a browser that does not display web pages that contain certain forbidden keywords is less protective of free speech than one who does display such pages.

A further relevant distinction, mirroring the distinction between normative and descriptive ethics, is that between a *normative* and a *descriptive axiology*. A descriptive axiology of technology or culture merely analyzes implicit values and norms, whereas a normative axiology utilizes a certain value system or thick conception of the good to critique particular value implications of technology or culture. For example, a normative axiology of video games could analyze them as embodying or promoting hedonism and weakening community (which would be a descriptive axiology), and subsequently fault them for this from a perfectionist and communitarian point of view (normative axiology). In addition to a need for axiological studies of technology and its co-evolved cultural forms, there is also a need for an axiological analysis of attitudes to and critiques of technology and technological culture in public and academic discourse.

An axiology of technology appraisals by comprehensive doctrines would study the value judgments and value discourses by representatives of comprehensive doctrines in their response to new technologies and their co-evolved cultural forms. For example, an axiology of protestant-Christian responses to the Internet and its culture would analyze the value judgments of representatives of this religious tradition in various writings and discourses. A extensive analysis of this sort may end up involving a detailed study of the doctrine's beliefs about and attitudes to

other cultural phenomena and artifacts to which the Internet is related: attitudes to and value judgments about modernity, the city, popular culture, technology in general, liberalism, and so on. Axiologies may also be performed of technology critiques of independent critics whose views are not tied to a comprehensive doctrine. Such axiologies would lay out and criticize the implicit value assumptions and conceptions of the good in a critique. Axiologies of technology critiques may be either descriptive, merely analyzing values implicit in these critiques, normative, critiquing these values from the point of view of a particular conception of the good, or *critical*, challenging the internal consistency of a critique and the validity of a its empirical assumptions regarding the relation between means and valued ends.

Implications for the Cultural Analysis of New Media

By means of illustration, I will now discuss four controversies in the cultural assessment of new media and will attempt to show how an axiological analysis can be of use in clarifying and critiquing the assumptions, arguments and presuppositions in these controversies.

Virtual reality and hedonism: The emergence of virtual reality technology has yielded a situation in which Nozick's "experience machine" is no longer an idea in a thought experiment but is becoming a real technology that tests our beliefs about the good life. While truly immersive VR still has technological limitations, the reality is that hundreds of millions children and adults spend a large part of their waking lives playing video games, which constitute a less immersive but still absorbing kind of virtual reality. Are these people wasting their lives or are they instead living the good life? Albert Borgmann's claim that virtual reality and cyberspace presents an illusory escape into another reality can be contrasted with Philip Zhai's claim to the effect that we should not be afraid to embrace Nozick's experience machine. Zhai presents an extended argument that one could recreate the whole empirical world in virtual reality, and that the distinction between such a world and the real world is no longer meaningful (Zhai, 1998). Yet, in spite of Zhai's best effort to create a metaphysical argument for his position, it will be clear that any choice for or against living a large part of one's life in virtual reality will depend on precisely one's attitude towards hedonism: is pleasure one's highest good or does one's well-being also depend on the veracity of one's pleasurable experiences? A hedonist reply can be contrasted with Albert Borgmann's objective, Aristotelian account of well-being, which commits him to deny that such virtual experiences can have great worth.

The instrumental value of cyberspace: An assessment of the instrumental value of a technology in relation to one's conception of the good may be difficult, because the meaning and consequences of (new) technologies may be ambiguous and opaque. In a critique of Borgmann's critical stance to cyberspace, Peter Paul Verbeek has argued that Borgmann wrongly holds that cyberspace offers us a substitute for reality that, in Borgmann's words, has cast a "lamentable pallor" on reality (Borgmann, 1999). Verbeek here does not attempt to counter Borgmann's Aristotelian account of well-being, but merely argues that Borgmann's negative assessment of the instrumental value of cyberspace for Aristotelian *eudaimonia* is false. According to Verbeek, cyberspace does not so much create an alternate reality as mediate existing reality, and can for this reason be as engaging. Borgmann (2002) responds that he holds that cyberspace both mediates reality, which he rates positively, and substitutes for it, which he rates negatively. He concedes, however, that the preponderance between these two uses will depends on actual uses of the technology, and acknowledges the relevance of social science data to settle this point.

Virtual communities and conditions for well-being: The debate on whether virtual communities

can serve as good substitutes for geographically localized communities is another debate that can be understood better by using axiological concepts and theories. In these debates, on can find two kinds of disagreement: disagreements about intrinsic goods and disagreements about instrumental goods. Disagreements about intrinsic goods concern the necessary ingredients for well-being. Communitarian critics of virtual communities, like Langdon Winner, have argued that strong community ties are an important ingredient for well-being, that such ties are not realized in most virtual communities, and that virtual communities negatively affect the formation and maintenance of geographical communities. Proponents of virtual communities have either denied the first or the second claim, and have either denied or ignored the third claim. Here, again, we see disagreements about intrinsic value as well as about instrumental value.

The Internet and Orthodox Judaism: In 2000, a group of leading orthodox rabbis in Israel, the Council of Torah Sages, issued a ruling banning the internet from Jewish homes, claiming that it is "1,000 times more dangerous than television" (which they banned thirty years earlier). The ruling required that all persons not given permission for Internet use by the Council to delete the Internet browser from their Windows program. The Council described the Internet as "the world's leading cause of temptation" and "a deadly poison which burns souls" that "incites and encourages sin and abomination of the worst kind." The Council explained that it recognized benefits in the Internet, but saw no way of balancing these with the potential cost, which they defined as exposure to "moral pollution" and possible addiction to Internet use that could quash the motivation to learn Torah, especially among children.[5] Using the framework that has been developed here, this ruling can be analyzed as a defensive action by leading figures in a comprehensive doctrine, a variety of orthodox Judaism called Hareidi, aimed at preserving the central values of this doctrine, including the highest good, which is obedience to God's law as laid out in the Torah. These leading figures perceived both instrumental benefits and harms in the Internet, relative to their doctrine's conception of the good, went on to conclude that the harms were greater than the benefits, and concluded that it was not possible to make changes in the technology nor adaptations in the practices and norms of their doctrine to preserve these benefits while minimizing the harms, leading them to the strong sanction of prohibiting Internet use.

4. Conclusion

It was argued in this essay that an important class of normative and evaluative analyses of new media cannot be classified as either belonging to ethics, political theory, or aesthetics, and that while these critiques concern cultural aspects of new media, this is not a distinguishing feature of them, because cultural critiques can also be political, ethical or aesthetic. It was argued that these analyses are characterized because they address our general idea of the good. It was then argued that these analyses could be usefully related to philosophical theories of the good and theories of value, of which an account was subsequently given. This account was then applied to the notion of culture to develop the idea of a theory of the culturally good, or cultural quality, and further applied to technology to develop the concept of an axiological study and critique of technology and technological culture. The axiological study of new media, new media culture, and appraisals of new media (culture) was presented as a specific variety of such studies of technology, with special importance because of the profound cultural transformations that have accompanied the diffusion of new media. It was argued that axiological analyses of new media and their appraisals can help clarify current debates on new media, and can help in the development of better critiques of new media (culture). Some example analyses were given to support this claim. Obviously, the present account is still sketchy and programmatic, but it has

[5] *Ha'aretz,* January 7, 2000.

both a solid basis in moral theory and philosophy and in science and technology studies, and seems to be helpful in analyzing, clarifying and critiquing issues in new media and new media culture. It is to be hoped, then, that this account can be developed further for the philosophical study of new media and technology at large.

References

Ben-Ze'ev, A. 2004. *Love Online. Emotions on the Internet*. Cambridge University Press.

Borgmann, A. 1999. *Holding On to Reality: The Nature of Information at the Turn of the Millennium*. University of Chicago Press.

Borgmann, A. 2002. "Response to My Readers," *Techné: Journal of the Society for Philosophy and Technology*, 6(1): 110-125.

Brey, P. 1997. "Social Constructivism for Philosophers of Technology: A Shopper's Guide," *Techné: Society for Philosophy and Technology,* 2(3-4): 56-78.

Brey, P. 1998. "New Media and the Quality of Life," *Techné: Society for Philosophy and Technology,* 3(1): 1-23.

Brey, P. 2000. "Method in Computer Ethics: Towards a Multi-Level Interdisciplinary Approach," *Ethics and Information Technology,* 2(3): 1-5.

Brey, P. 2001. "Disclosive Computer Ethics," in R. Spinello and H. Tavani, eds., *Readings in Cyberethics*, Sudbury, MA: Jones and Bartlett, 51-62.

Brey, P. 2003. "The Social Ontology of Virtual Environments," in D. Koepsell and L. Moss, eds., *John Searle's Ideas About Social Reality: Extensions, Criticisms and Reconstructions.* Blackwell Publishing.

Brey, P. 2005. "Artifacts as Social Agents," in H. Harbers, ed., *Inside the Politics of Technology*. Amsterdam University Press.

Bijker, W., Pinch, T. and Hughes, T., eds. 1987. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. MIT Press.

Castells, M. 1996. *The Information Age: Economy, Society and Culture. Volume 1. The Rise of the Network Society.* Oxford: Blackwell.

Carson. T. 2000. *Value and the Good Life*. Notre Dame University Press.

Chaney, D. 1996. *Lifestyles*. Routledge.

Dreyfus, H. 1999. "Anonymity versus Commitment: the Dangers of Education on the Internet," *Ethics and Information Technology,* 1: 15-21.

Dyson, K. and Homolka, W., eds. 1996. *Culture First! Promoting Standards in the New Media Age.* Cassel.

Etzioni, A. 1998. "Voluntary Simplicity: Characterization, Select Psychological Implications, and Societal Consequences," *Journal of Economic Psychology,* 19: 619-43.

Featherstone, M. 1991. *Consumer Culture and Postmodernism*. London: Sage.

Floridi, L. 2002. "On the Intrinsic Value of Information Objects and the Infosphere," *Ethics and Information Technology,* 4 (4): 287 - 304.

Fountain, J. 2001. *Building the Virtual State: Information Technology and Institutional Change.* Brookings Institution Press.

Gumbrecht, H. and Marrinan, M., eds. 2002. *Mapping Benjamin: The Work of Art in the Digital Age.* Stanford University Press.

Hill, K. and Hughes, J. 1998. *Cyberpolitics. Citizen Activism in the Age of the Internet.* Rowman and Littlefield Publishers.

Hurka, T. 1993. *Perfectionism.* New York: Oxford University Press.

Inglehart, R. 1990. *Culture Shift in Advanced Industrial Society*. Princeton, NJ: Princeton

University Press.
Inglehart, R. 1997. *Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies*. Princeton, NJ: Princeton University Press.
Johnson, D. 1997. "Is the Global Information Infrastructure a Democratic Technology?," *Computers & Society,* 27: 20-26.
Johnson, D. 2000. *Computer Ethics*, 3rd ed. Upper Saddle River: Prentice Hall.
Larmore, C. 1996. "The Right and the Good," in *The Morals of Modernity*. Cambridge University Press.
Lievrouw, L. & Livingstone. S., eds. 2002. *Handbook of New Media: Social Shaping and Consequences of ICTs.* Sage.
Lister, M., Dovey, J., Giddings, S., Grant, I. and Kell, K. 2003. *New Media. A Critical Introduction.* London and New York: Routledge.
MacKenzie, D., and Wajcman, J., eds. 1999. *The Social Shaping of Technology, 2*nd ed. Buckingham: Open University Press.
Malinowski, B. 1944. *A Scientific Theory of Culture and Other Essays.* Chapel Hill: University of North Carolina.
McLuhan, M. 1964. *Understanding Media.* McGraw Hill.
Merton, R. 1957. *Social Theory and Social Structure*, revised and enlarged. London: The Free Press of Glencoe.
Misa, T., Brey, P. and Feenberg, A. 2003. *Modernity and Technology*. MIT Press.
Mitchell, W. 1995. *City of Bits: Space, Place, and the Infobahn.* Cambridge, MA: MIT Press.
Mossberger, K., Tolbert, C. and Stansbury, M. 2003. *Virtual Inequality: Beyond the Digital Divide.* Georgetown University Press.
Nissenbaum, H. 1998. "Values in the Design of Computer Systems," *Computers in Society*, 38-39.
Nussbaum, M. 1986. *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy.* Cambridge University Press.
Nozick, R. 1974. *Anarchy, State, and Utopia*. Oxford: Basil Blackwell.
Parfit, D. 1986. *Reasons and Persons.* Oxford University Press.
Pfaffenberger, B. 1992. "Technological Dramas," *Science, Technology and Human Values,* 17: 282-312.
Rescher, N. 2004. *Value Matters: Studies in Axiology.* Frankfurt: Ontos Verlag.
Rawls, J. 1971. *A Theory of Justice.* Cambridge, MA: Harvard University Press.
Rawls, J. 1993. *Political Liberalism.* Columbia University Press.
Ross, W. 1930. *The Right and the Good.* Oxford University Press.
Saco, D. 2002. *Cybering Democracy: Public Space and the Internet.* University of Minnesota Press.
Sclove, R. 1995. *Democracy and Technology.* New York: Guilford Press.
Shaw, D. and Newholm, T. 2002. "Voluntary Simplicity and the Ethics of Consumption," *Psychology and Marketing,* 19 (2), 167-185.
Slater, D. 1997. *Consumer Culture and Modernity.* Polity Press.
Tavani, H. 2003. *Ethics and Technology: Ethical Issues in an Age of information and Communication Technology*. Wiley.
Turkle, S. 1995. *Life on the Screen: Identity in the Age of the Internet.* New York: Simon & Schuster.
Verbeek, P.P. 2002. "Devices of Engagement: On Borgmann's Philosophy of Information and Technology," *Techné: Journal of the Society for Philosophy and Technology,* 6(1): 69-92.
Virilio, P. 1994. *The Vision Machine.* Indiana University Press. [trans. from French, *La Machine*

De Vision, 1988]

Wardrip-Fruin, N. and Harrigon, P., eds. 2004. *New Media as Story, Performance, and Game.* MIT Press.

Wardrip-Fruin, N. and Montfort, N., ed. 2003. *The New Media Reader.* MIT Press.

Winner, L . 1980. "Do Artifacts Have Politics?," *Daedalus,* 109: 121-136.

Winner, L. 1997. "Cyberlibertarian Myths and the Prospects for Community,"*Computers and Society,* 27:3: 14-19.

Zhai, P. 1998. *Get Real. A Philosophical Adventure in Virtual Reality*. Rowman and Littlefield Publishers.

Theorizing Technological and Institutional Change: Alienability, Rivalry and Exclusion Cost

Paul Thompson
Professor of Philosophy
W. K. Kellogg Chair in Agricultural, Food and Community Ethics
Michigan State University

Abstract
Formal, informal and material institutions constitute the framework for human interaction and communicative practice. Three ideas from institutional theory are particularly relevant to technical change. Exclusion cost refers to the effort that must be expended to prevent others from usurping or interfering in one's use or disposal of a given good or resource. Alienability refers to the ability to tangibly extricate a good or resource from one setting, making it available for exchange relations. Rivalry refers to the degree and character of compatibility in various uses for goods. The paper closes with a note on how attention to these factors might be useful ways to conceptualize what Langdon Winner has called "the technological constitution of society," and what Andrew Feenberg has theorized as "secondary rationalization," as well as within more practical contexts of technical research, development and design.

Keywords: industrialization, capitalism, institutional economics, biotechnology

Philosophy has long been concerned with the nature, rationale and legitimation of formal institutions such as law, education and social bureaucracy, and has traditionally reflected on informal institutions in the realm of culture, habit and tradition. Yet the plasticity of the manner in which material reality also frames human interaction has often escaped philosophical inquiry. 20th century social science developed penetrating analyses of formal and informal institutions on many levels, yet like philosophers, social scientists have neglected the implications of their ideas for the transformation of the material world. To contextualize this theoretical gap, I begin by retelling a familiar story of modernization in succinct form as a story of institutional change and then shift abruptly to an equally succinct discussion the three analytic concepts that appear in the title: alienability, rivalry and exclusion cost. I will briefly discuss how philosophical evaluation of changes in formal or informal institutions has centered on one or more of these factors, while also offering examples of technical change where changes in exclusion cost, alienability and rivalry restructure human relationships in very similar ways. After these stage setting exercises, we arrive at the main philosophical task: to merge these concepts into our explanatory framework for industrialization, technical change, the growth of capitalism and the emergence of the modern era. The concepts of alienability, rivalry and exclusion cost and the theoretical framework of institutional change allow us to pose questions that have been asked by Herbert Marcuse, Langdon Winner and Andrew Feenberg in a new way: If technology is in part responsible for the shape of our institutions, and if institutional change in the sphere of law and custom can be subjected to philosophical critique and democratic guidance, why shouldn't technology be subjected to the same critique and guidance? A more pointed form of the question can be posed to scientists and engineers at work in technical innovations: why shouldn't technical designers account for factors such as exclusion cost, alienability and rivalry in considering alternative

designs? Why shouldn't the developers of technology be socially and politically accountable for consequences accruing from alterations in alienability, rivalry or exclusion cost, as well?

The Political Economy of Institutional Change

Philosophers and other social theorists have long focused on underlying structures or patterns of social organization, attempting to understand both the mechanisms and implications of change within them. But these structures and patterns have been thematized in almost innumerably many different ways. I want to focus on patterns and transitions associated with transformations that have been characterized as rationalization, commodification and institutional change. It should be obvious why the philosophy of technology should take an interest in these transformations, for they are closely associated with industrialization, the rise of capitalism and with various theses of technological determinism. The diffuse body of theory that has been associated with these transformations can be summarized for present purposes as emphasizing the transition from informal to formal institutions.

Institutions are standing practices or patterns of human activity that can be described in terms of rule-governed behavior. *Formal* institutions are those that are explicitly articulated as rules, and that are reproduced and enforced by organized social entities, especially the state. Hence, formal institutions are laws and public policies. *Informal* institutions are standing practices that subsist on the basis of common knowledge, tradition and culture. They are reproduced through legend, lore, apprenticeship, imitation and perhaps all manner of common experience. Their enforcement mechanisms can include approbation, praise, shunning or group inclusion but consist mainly in the way that they constitute the framework for successfully negotiating the most basic tasks in social life (Commons, 1931). Although vague, this simple set of definitions provides a basis for interpreting the last millennium of European history as the gradual displacement of informal institutions by formal regimes of law and policy.

Philosophers of the Enlightenment and early Modern Age were deeply complicit in this displacement, typically viewing formal institutions as superior in virtue of their capacity for explicit articulation, widespread application and critical evaluation. A rule that cannot be clearly stated cannot be criticized or justified, much less enacted by a civil authority, even if it can be reliably followed by those who are appropriately socialized. Thus philosophers' predilection for argument, demonstration and verbal disputation disposed them to regard formal institutions as inherently rational. Or perhaps we should say, as C. B. MacPherson (1962) did, that those interests most consonant with the evolution of property rights and state authority naturally aligned themselves with philosophers who were advocating explicit, rational evaluation of society's rules. But the philosophical bias in favor of formal institutions began to wane in the Romantic period, as the new wave of philosophy begins to pine for a lost sense of belonging and community solidarity. In 1897 the German sociologist Ferdinand Tonnies (1855-1936) theorized modernization as a transition from *Gemeinschaft* to *Gesellschaft*, and in 1914 Max Weber (1864-1920) characterized it as a process of rationalization toward increasingly bureaucratic decision making.

For Karl Marx (1818-1883) and subsequent Marxists, the alienation or estrangement of labor is a turning point in this long process. Marx's *1844 Manuscripts* explore the metaphysical and moral significance of this event, but in what, exactly, does the alienation of labor consist? Economic historian Karl Polanyi (1886-1964) described it as a series of legal and policy changes by which manorial social relations give way to capitalist relations. Under traditional social relations,

peasant labor was bonded to a particular parish or parcel of land. By common consent (or at least accepted practice), laborers were both minimally maintained by the liege lord, but also unable to work for compensation beyond the parish borders, except by the express permission of parish authorities. This system stifled the development of the factory system, which demanded large numbers of laborers at specific locations. It was abandoned in favor of a system of wage labor that, just as John Locke (1634-1704) had argued, made each individual the owner of their own labor, but which also obligated them to sell it at the going rate in order to obtain subsistence. These legal and policy changes thus allowed labor to be alienated both from the soil and from the social relations in which it had previously been embedded, and to be sold as a commodity good on a competitive market (Polanyi, 1944).

British labor historian E. P. Thompson (1924-1993) argued that, in fact, a more extensive set of transformations had contributed to the making of a working class, transformations that predated the industrial age by centuries. These included the alienation of ordinary food from the circumstances in which the production, distribution and consumption of grain had been embedded so that it could be traded as a commodity good. Before modernization, the grain growing in an English field would have been considered the common property of the parish. An elaborate system of informal concessions governed the share to which each parishioner (not to mention the lord) was entitled, as well as the tasks such as harvesting, milling, or baking that each was obligated to perform. Although this system might be theorized as a regime of exchange in which goods and services are traded at fixed prices, Thompson analyzes it as a "moral economy" governed more firmly by mutual expectations than by formal institutions of ownership and regulated exchange. The system was gradually monetized during the early stages of modernization, with entitlements becoming defined as forms of income and many exchanges taking the form of cash sale. As roads and wagons improved, the farmers who harvested and bagged grain (not to mention the lord) saw opportunities to sell it in other villages or wherever prices were best, ignoring the informal expectations (the assessments and shares) that governed the distribution of grain under traditional practice. How are we to interpret this situation? Do the farmers have a right to seek the best price for their grain, or is it the common property of the village?

Natural law philosophy tended to notice a few key things about grain. First, the farmers who come into first possession of a parcel of grain through the labor of sowing and harvesting can easily keep tabs over its location and use, and it is fairly easy for the grain to change hands by sale or gift. Furthermore, once consumed for one use, the grain is gone. It cannot be re-eaten by another. These natural characteristics of grain were seized upon by natural law theorists, who saw a sack of grain as something naturally fit for property rights. Thus, the natural law theorists endorsed the of farmers' right to claim ownership of the grain, and redefined the sack of grain as a commodity good, replacing the informal moral economy with the formal institution of state sanctioned commodity exchange, (Thompson, 1971). Thus did Marxist theoreticians theorize the transition from informal to formal norms as one of commodification where economic practices of production and distribution are disembedded from more complex social relations and become available for monetized exchange. Thus also did they theorize political economy as a tool for capitalism and commodification.

One lesson to take from this attenuated overview of social history is the emphasis that is placed on the decline of informal institutions and the rise of formal ones. The theoretical focus is on the creation of a social apparatus that formulates and enforces the principles according to which human activity is to be guided. Much attention is given to state actors, but non-state

bureaucracies (such as the Dutch East India Company) are active in more detailed accounts of the transition, and they become more and more active as laws of incorporation become common. A second lesson is that the capacity for rational rule-governance, as well as for rational revision of rules, depends upon the recognition of social relations that can be disembedded from the thick practices of common custom. Thus if institutions and their transformation are to be made into a subject for philosophical deliberation or public choice, there is an implicit bias against customs and traditions that emerge through evolutionary or adaptive social processes. As we shall see, this carries over into a bias favoring the deliberative review of formal institutions instead of material practice. The third lesson is that the process of disembedding often involves the creation of alienable goods, goods whose production and distribution can be controlled. This process centers on altering the *alienability,* the *exclusion cost* and the *rivalry* of goods.

Alienability, Rivalry and Exclusion Cost

Until fairly recently, neo-classical economic theory assumed that a rational person would always exchange a good "A" for a good "B" whenever the person preferred having "B" over "A". This assumption had long been recognized as exceedingly unrealistic in virtue of the fact that circumstances of the exchange could override the preference for "B" over "A". The individual would have to know that the opportunity for exchange was available, for example, and the greater value of "B" would have to be sufficient to make it worthwhile for the person to take the trouble to make a trade. Furthermore, in the real world, trading "A" for "B" sometimes means that one also has to accept "C", as anyone who has ever purchased a puppy can attest: cuddles and endearing looks come bundled with training responsibilities and interruptions in the dead of night. This extra baggage can make the whole package seem less attractive than it otherwise might. Such circumstances have been theoretically characterized as "transaction costs," by new institutional economists, who have made numerous strides to make economic theory more realistic. Alienability, rivalry and exclusion cost are three among many factors that have been analyzed as contributing to transaction cost. For the most part, institutional economists have not abandoned the neo-classical assumption that rational behavior is always concerned with economizing, and as such, they have tended to think that reducing transaction costs is always a good thing (North, 1990). Although I do not share these framing assumptions, I will borrow heavily from the institutionalists' characterization of alienability, rivalry and exclusion cost in order to make my own theoretical points.

Alienability is the degree to which a good or potential item of use can be extricated from one setting or circumstance so that it can be transported to or utilized in another. A critical aspect of alienability is the ease with which something in the possession or employ of one human being can be transferred to the possession or employ of a different human being. The right to life is characterized as an inalienable right because life can only be lived by specific individuals, it can't be given or sold to someone else. Hence the *right* to live can only be exercised by the person whose life is at stake, it cannot be alienated from that person and exercised by someone else. Alienability is in this sense a metaphysical characteristic of goods that determines whether the goods can be meaningfully subject to exchange. Alienability is a necessary prerequisite for any item of property, at least as this notion has been understood in the natural law tradition. Most analyses of alienability focus on formal legal institutions rather than metaphysics, and the question is whether it is legally permissible to alienate a good (often labor) and to offer it for exchange. But since laws can change, legal alienability can change. It is situational rather than metaphysical. Both legal and metaphysical alienability may seem to be absolute: something is alienable or it is not. But an institutional focus shows that alienability can come in degrees.

Making it easier to "unbundle" goods—to alienate one good from another—affects transaction cost, and dramatically affects the price. "I will take your puppy for $100 if you agree to supervise the housebreaking, but I will only give you $10 if I must do it myself." Thus, in addition to pure metaphysical alienability (something that is just not the *kind* of thing that can be alienated) and pure legal alienability (it's legal to alienate that thing or aspect or it's not), there is a relative and negotiable domain in which the cost of alienating the good is reflected in whether the good is typically exchanged or not.

It is important to note, however, that a fairly large component of sociability depends on the degree to which various items or goods are alienable or in fact alienated from one another. For Thompson's peasants, the fact that it was rather difficult to separate large quantities of grain from inland locales where it was grown prior to the advent of canals, better roads and boats or wagons made for a situation conducive to the embedded relations of production and exchange that were characteristic of manorial society. Here, the inalienability of grain from place was a situational rather than a metaphysical necessity, or even a legal practice. Farmers and lords may have had a legal right to sell grain but they were very limited in who they could sell that grain to. Other situational forms of inalienability include the impossibility of separating a musical or theatrical performance from the person of the artist prior to the invention of photography and audio recording. A sixteenth century minstrel might have had the legal right to sell the right to enjoy his performance of a song to someone who was not physically present and able to hear it in person, but this is not a right that would have occurred to anyone, much less had much cash value. After Edison, the right becomes meaningful. Prior to the legal reforms documented by Polanyi it was also legally impossible to separate the labor power of a worker from the parish in which he had been born.

These situational types of inalienability can be changed, in the latter case by changing the law and in the former cases through material transformation. But we may speculate that in virtually every case it is difficult to imagine how goods might be alienated one from another until it has become obvious that it can be done. In our own time, traits that might have been thought to be inalienable characteristics of certain plants or animals can now be readily encoded in genetic sequences and transferred to totally different plants and animals through genetic engineering. These traits (or at least the genes that confer them) have even been alienated from organisms altogether and put on the market all by themselves in the form of licenses that plant or animal breeders may purchase so that they may then transfer the trait to different organisms. It would have been difficult to conceptualize the growth rate of a fish as something that could have been alienated from the species or type of fish prior to this development in genetics. If you want fast growing fish, you would have to get fish that grow quickly. But growth rate has now been alienated and it is now possible to build a fast growing fish (or a fast growing anything) simply by buying the gene construct (Muir, 2004).

Rival use or *rivalry* is the degree to which alternative goods or uses of goods come into competition with one another. One way in which two alternative uses of a good can compete with one another is when they are consumed in use. Eating the grain is a comparatively rival use because it can only be eaten once, and this use exhausts the possibility of its being used by another person or in another way. Enjoying the scenic beauty of the waving fields of grain is a non-rival use because not only can more than one person obtain this good from a single field of grain, scenic beauty can be enjoyed again and again. Economists also use the concept of rivalry to describe the relationship between two or more goods that can be substituted for one another and which therefore come into competition with one another in market relations. Thus beans and corn

may be rival in that both can be eaten, and food shoppers may opt for beans when the corn is too expensive. But beans and corn are non-rival in other markets: you can't use beans to make Tennessee whiskey, so a moonshiner is never in the market for beans.

Rivalry is thus situational, and situations can change. Since antiquity, farmers have made use of seeds by saving a few from each year's crop and planting them in the following spring to grow another crop. This year's crop of corn or beans produces food, but some of the corn and beans that could be eaten can be used as seed, which can be planted again. In this sense, using a seed to plant a crop is a qualified non-rival use. It does not deplete the amount of the good available for future uses, though it does make the good temporarily unavailable while the crop is in the ground. Genetic use restriction technologies (GURTs), or so-called "Terminator" genes, can be used to create seeds that when sown as a crop will not produce more seeds. Although the corn or beans from a GURT crop can be eaten, if a farmer saves them to plant, she will be sorely disappointed for they are infertile and cannot function as seeds. GURTs thus transform the use of seeds to sow a crop from a non-rival to a rival use (Conway, 2000).

Alienability and rivalry are critical to the creation of exchange relations because they influence the degree to which a good is amenable to the process of and the need for exchange. Goods that cannot be alienated one from another effectively become a single good for the purposes of exchange, if they can be exchanged at all. Rival goods are depleted by use, and hence must be obtained and replenished prior to any use, or they may substitute for one another, also affecting the need to obtain them through exchange. Thus, whether exchange takes the form of sale, gift or grant, it is primarily alienable and rival goods that are the object of exchange. Or to put this in somewhat different terms, although human beings can exchange glances, insults and affection, it is the exchange of alienable and rival goods such as a sack of grain, a team of oxen or a day's work in the fields that constitute the paradigmatic form of the economic social relationship.

The degree to which alienable and rival goods precipitate social relations characterized by commercial exchange also depend on the ease with which the various uses of a good can be limited or controlled through access or possession. *Exclusion cost* is the outlay in time, trouble and expenditure of resources that is required in order to prevent others from having access to a particular good or item of property. Like alienability, exclusion costs are in large measure a function of the material characteristics of the goods human beings utilize and on which they rely. Oxygen and vitamin D are alienable and rival goods, but it is fairly difficult to prevent people from having access to air and sunshine. It is, in contrast, fairly easy to keep jewels and trinkets where no one else can get them hence the latter have more typically been understood as saleable items than the former. Items with very high exclusion cost are unlikely to be traded commercially.

Like alienability and rivalry, exclusion cost is amenable to situational variation. Situational change in exclusion cost has often taken the form of material manipulation of either the goods in question of the circumstances in which they reside. Locks and fences are the classic technologies of exclusion, and a better lock will lower the cost of excluding others every time. It has also been possible to reduce exclusion costs through the development of informal institutions. Simply declaring that certain parties have an exclusive right to use a good will suffice in many cases. Queuing for service is among the most venerable of informal institutions in Western cultures, and everyone recognizes that the person at the front of the line has an exclusive right to be served next. If being served next is the good in question, we may thus say that for the first in the queue, the cost of excluding anyone else from this good is very low. By common consent, customary

recognition of this right saves everyone a load of time and trouble, making the cost of many daily transactions far more reasonable.

When customary rights of exclusion are threatened, it is always possible to bring in the coercive power of the state to back them up. The police represent a formidable way of lowering exclusion cost for all manner of private property. A person who would have to guard or defend an item of property can call on the police to do it, and the knowledge that arrest and prison are among the possible consequences of an unlawful taking raise the cost of theft, simultaneously lowering the cost of exclusion. Copyright and patent laws represent formal institutions that place the coercive power of the state behind a broad array of exclusive practices, even when no tangible property exists. The legal remedies of intellectual property law vastly reduce the cost of preventing others from using one's intellectual creations through intimidation, bullying, spying and other forms of self help.

Alienability, rivalry and exclusion cost represent features of the various items and entities in the world, including personal services as well as material things, that collectively determine which items and entities come to be the object of exchange relations, and which ones remain embedded within a more inchoate and presumptive context of social practice. It is very likely that anything alienable, rival and excludable will be regarded as an item of personal or private property. It should not be surprising that when goods are lacking in one or another of these three dimensions, a few people try make up for it either by passing laws or by changing the world in a material way. As institutional economists have developed their analysis of these traits, they brought the economists' bias that enabling transaction is always a good thing. They also bring the social scientist's bias of focusing on social practice, and especially on formal institutions. As such, they have tended to focus on legal or policy reforms that would lower transaction costs. But as my illustrations demonstrate, it is equally possible to affect alienability, rivalry and exclusion cost with a technical as with a legal change.

Technology, Social Practice and Political Change

Now it is time for a few observations that may seem profound if they do not seem altogether obvious. First, a fair proportion of internal political conflict over the last millennium has either involved or been precipitated by changes in the alienability, rivalry or exclusion cost of goods. State-led efforts to rationalize embedded activities of production, distribution and consumption by enacting laws that create formal institutions for exchange are at the bottom of social critiques offered by Marx, Polanyi and a host of other social theorists. For example, in *Wage Labor and Capital,* the *1844 Manuscripts,* and *Das Capital,* Marx challenges the viability of the institution of wage labor on various grounds, sometimes stressing the moral plight of the wage laborer, other times arguing that the social prerequisites for the reproduction of the labor force were simply not met by the institution as it had taken shape in 19th century Europe. But the institution of wage labor was a function of legal changes that had altered the alienability of labor power in two senses that are not clearly articulated among the four that Marx mentions in his famous analysis. First, the laws and customs that had tied labor power to land were eliminated, allowing labor power to be alienated from a specific geographical locale, and hence also the social setting in which it had theretofore been embedded. Second, labor power had previously been a bundled good, thoroughly entangled in the person of the laborer and not to be had without also accepting at least minimal responsibilities to sustain the person.

The first of these alterations in the alienability of labor power is a knife to the heart of *Gemeinschaft*, the intense local sociability that we perhaps nostalgically associate with the pre-industrial world, while the second is the source of most left-leaning complaints against capitalism. It was, of course, also possible to see this change as progressive in virtue of the way that labor markets allocate labor power to society's most valued use. The use of formal institutions to change the alienability of labor power thus lies at the core of social theories such as Tonnies' that stress industrial society's loss of community solidarity, socialist theories that stress capitalism's inability to meet the basic needs of the poor, and neo-liberal theories that stress the compensating benefits of industrial growth. This is all old news, of course, but what remains striking in the social theories of industrialization is the bias toward formal institutions. For material changes in alienability, rivalry and exclusion cost are every bit as important in creating the watershed transformations that led to the industrial world.

To take one example, labor power that is highly specialized is comparatively non-rival. To be sure, to the extent that labor is a function of time spent working, all labor is highly rival, because nothing is more thoroughly depleted by use than time. However, work that requires a lot of skill or special training can be done by many fewer workers in the pool. Thus, the deskilling often associated with machines and assembly line operations converts labor power into a more rival good. Work that can be done by almost anyone creates a labor market in which many more workers compete for jobs, driving down wages. Deborah Fink's study of late 20th century meatpacking shows how packing companies introduced new technologies requiring considerably less skill precisely as a union-busting tactic that redefined work rules and brought a new group of unskilled (mostly immigrant and female) workers into the workforce, (Fink, 1998). If low-skill, low-wage workers are able to perform work once done by those who have the skills, strength and stamina needed for traditional meat cutting, there are more rivals (more types of labor) that substitute for one another from the employer's perspective. Such materially and technologically based changes in labor needs for manufacturing are emblematic of industrialization.

For a second example, let's return to E. P. Thompson's peasants, who rioted when local farmers asserted that their right to sell grain in a neighboring village was in fact a right to seek the best price in more extensive commodity market created by expanded modes of transport. Here what had once been assumed to be community property, if not by legal right then by the informal norms of the "moral economy," became a more readily and hence more thoroughly alienable good, protected by private property rights and available for sale to the highest bidder. Although grain itself was not changed in this transition, as it has been in the case of Terminator seed, what was changed was the material infrastructure—wagons and roads—and it was this technological change that made grain into a good that was practical to alienate from the local community for the first time. As noted above, these transformations preceded the period of industrialization by several hundred years, but they contributed to the process we know as modernization as surely as did the creation of a factory system.

Much ink continues to be spilled over industrialization, modernization, capitalism and technological determinism, and the analysis (not to mention the examples) that has just been given cannot be disentangled from the raging debates over how and whether these things fit together or don't. Tom Misa, Phillip Brey and Andrew Feenberg have published a collection of essays by multiple authors which examine the tensions that animate these debates through a number of different theoretical and disciplinary perspectives. The main thrust of most essays is that modernization theory and empirical studies of technology are passing like ships in the night, and that more focused attention on the gap between these two bodies of scholarship would be a

good thing (Misa, Brey and Feenberg, 2003). Although the argument thus far has drawn upon the literature of modernization and the debate over technology and the engines of history to create a philosophical context, my goals are not to take sides in that debate as much as to fall in line with Misa,. Brey and Feenberg's call for a kind of theory that would fill in the gap. Thus, the short version of the long story of how we got where we are today will now be set aside (at least for awhile) in pursuit of new theoretical goals.

Technology: An Institutional Approach

We may thus focus on three modes of transformation for the institutional infrastructure of society. The first of these is formal and reflects the processes of bureaucratic decision making that were the focus of Weber's sociology. Institutions reflect the rules of the game for social interaction. Legislation, the courts and the administrative agencies of government each bring to bear various rule-governed procedures for revising those rules. The second mode is simply cultural change, the gradual transition in expectations, shared beliefs, custom and tradition that supports a vast array of informal institutions, most of which, like the clothing on our backs, fail to be particularly evident to us at the very moment that we participate in their social production and reproduction. Finally, there is technical change, the intentional modification and manipulation of the material world. Technical change shares an element of mindful deliberateness with formal institutional change. Technical changes, in other words, come about because some person or group intend for them to happen. Yet technical changes are often taken up gradually, with numerous adaptations and modifications that Andrew Pickering calls "tuning," (Pickering, 2005), and in this sense they share an apparently haphazard and evolutionary modality with cultural change.

Although it has long been obvious that technical change has a critical role in shaping history, it is perhaps still not widely accepted that some types of technical change also operate in the modality of institutional change. Part of the reason for this is that institutions seem to have a normative character that material objects do not have. Institutions are rules about what people are permitted to do. The institution of queuing for service is only effective because people think that they ought to behave as the institution demands. Take away this "ought" and you take away the institution. Ethics and political theory are normative discourses that attempt to state what *people* ought to do in given circumstances. There are no normative theories that attempt to state what things ought to do under any circumstances. Things are notoriously uncooperative when it comes to philosophical persuasion. Most people are inclined to think that they lack the capacity to follow normative advice in the fashion that philosophers have been most inclined to give it. The fact that many of our students also seem to lack this capacity has not persuaded philosophers to think that the problem might lie in the way that normative theories are articulated. Despite Bruno Latour's efforts to persuade us otherwise, philosophy as a discipline continues to insist that norms for things are a waste of time because things do not have minds and are incapable of intentional action.

It may be difficult to see things in the world as having any institutional significance at all simply because we do not, in the age of disenchantment, understand the material world as able to support a normative dimension. But there will not be institutions forbidding actions that are physically impossible. We do not, for example, have institutions that dictate when it is and is not appropriate to become invisible. Yet our need for decorum and privacy would surely have led people to form customs governing the practice of disappearing from view while remaining present as an observer if this were a capacity that people actually had. Similarly, although property rights, work rules and a host of other social institutions specify norms for the alienation of goods, for rival use and

for the right to exclude others from access to goods and services, it would be rather surprising if there these institutions did not closely track the material possibilities for alienation, rivalry and exclusion cost. Just beyond the domain of sheer metaphysical possibility there lies the socially crucial domain of cost. Here, the relative *ease* of alienation, controlling rivals and excluding others may be almost as determinative as metaphysical possibility in affecting whether we have formal or informal institutions. In places where it would be very difficult, that is, very expensive, to exclude others from access to sunshine, you can bet that there will be no informal norms (no rules) about whether or when it is appropriate to do so.

Furthermore, as long as material transformation of the world is comparatively minor or slowly paced, the process of adaptation and adjustment in social institutions that occurs in response to these changes will probably be absorbed into the background noise of ongoing cultural change. It is only when material changes result in relatively large changes in alienability, rivalry and exclusion cost that technical change can be distinguished from ongoing cultural change. Furthermore, it is only when such large scale changes become sufficiently frequent that it will become clear to people that technical change operates as a distinct modality of institutional change, as a class of human originated events having a patterned (if only vaguely predictable) impact on the texture and importance of human interaction. When this modality becomes clear, it will be obvious that even though things do not have minds, they do have normative implications. The material dimensions of alienability, rivalry and exclusion cost represent a "given" or natural infrastructure in which formal and informal institutions evolve, either by chance or by design. When those background conditions change, by chance or by design, the entire significance of social institutions can be altered.

Changing Things by Design

All of which raises the question, if changes in the formal institutions of society are appropriate targets for political philosophies and theories of justice, why not also the technological transformation of alienability, rivalry and exclusion cost? This is, I take it, a somewhat more focused restatement of a question that has been asked many times before. Herbert Marcuse's *One Dimensional Man* suggests that the failure to subject technical systems to normative scrutiny is both a political and a philosophical failure. The political failure resides in the increasing power of capital and commercial interests to dominate all forms of discourse in industrial society, while the philosophical failure consists in positivist doctrines that created an epistemological space in which questions about technical efficiency were regarded as "value free," (Marcuse, 1966). Today, philosophical positivism no longer maintains much influence over the practice of science and engineering, though its legacy no doubt lingers in the form of uncritical attitudes and institutionalized organizational practices that penetrate deeply into the social complex of technical innovation, development and regulation (Thompson, 2004). Resistance to Marcuse's demand for a critical philosophy of technology lingers, as well.

This lingering resistance may in part simply reflect the continuing influence of powerful economic interests, but Marcuse's characterization of technology has seemed too metaphysical, too Heideggarian, in fact too vague to provoke much critical reflection on the part of many. Langdon Winner has had more success in calling for critical evaluation of technology and technical change by describing what he calls "the technological constitution of society." This is a material and organizational infrastructure that predisposes a society toward particular forms of life and patterns of political response. Winner illustrates his idea with a number of examples, notably technological systems such as irrigation systems or electric power grids that dispose

societies toward centrally administered, hierarchical relationships of political power (Winner, 1986). We should notice that what in fact accounts for such tendencies is the way that these systems affect the alienability, rivalry and exclusion cost of the respective goods—water and energy—that they produce and distribute.

Centrally administered irrigation systems in the ancient world and contemporary electric power grids succeed in part because they represent technical solutions to real problems, but they also have the effect of converting goods that are comparatively non-rival and with high exclusion costs into goods that are just the opposite. Water and energy are virtually everywhere in most locales, though frequently not in large enough concentrations to accomplish certain critical tasks such as agriculture or manufacturing. In their natural state, however, water and energy have high exclusion costs; it takes a bit of trouble to keep people from having access to them. Natural water systems such as rivers and springs also serve a number of purposes simultaneously and in this sense are comparatively non-rival goods. Though generally depleted in use and in that sense naturally rival, energy in the form of wood and mineral fuels or localized wind and water mills is relatively specialized in the types of work it can be expected to perform. One type yields heat and the other mechanical power, and further technology is needed to reconfigure them for other purposes. Thus water and energy are relatively non-rival under these configurations of the material world, meaning, again, that the "markets" in which people access these goods will be distinct. The irrigation system and the power grid reduce exclusion cost as they increase rivalry, and the result is goods (i.e. water on tap or electrical energy at the wall outlet) that are far more amenable to centralized control *and* to commodity exchange than water and energy are without these technological infrastructures. What is more, both systems provide a way to alienate their respective goods from a local setting, much as wagons and roads transform the alienability of grain. Thus, alienability, rivalry and exclusion cost are part and parcel of what Winner has called the technological constitution of society. These traits, in fact, specify the politically important design parameters of a technological system more clearly.

Andrew Feenberg has been among the most recent theorists to call for the evaluation of technology in ethical and political terms. He has done so by arguing that technological systems undergo two phases of rationalization, one that might be characterized fairly positively in terms of technical parameters, and a second that has to do with the way that technological means and artifacts interface with networks of human actors. It is the second phase of interface that, in Feenberg's view, should be the focus of political and philosophical critique (Feenberg, 1999). But how can we characterize the boundary between humans and non-humans in a manner that allows us to bring traditional categories of political philosophy to bear? There are probably many ways to do this, some of which will clearly stress social parameters such as who stands to profit in terms of money or prestige when a given technology is widely adopted. Yet if technical systems rearrange the material world in ways that affect the alienability, rivalry and exclusion cost of goods, this will certainly impact the networks in which humans will be enrolled. Thus with Feenberg's secondary rationalization as with Winner's technological constitution, alienability, rivalry and exclusion represent ways to ask the philosophical and political questions in more pointed terms.

However, if the conceptual framework made available by institutional economics allows us to sharpen the questions we wish to direct at technology, it also results in a deflation of the thesis that technology needs to be questioned. First of all, it is clearly specific tools and techniques as utilized in specific situations that give rise to the material consequences I have been illustrating. We are not doing philosophy of technology in its woolliest, most metaphysical incarnation today.

Pragmatism is implicit in my general approach (see Hickman, 2001). Second, not all of these material changes will rise to the level of political importance. One would hardly object to better locks on the ground that they lower the exclusion costs for people who use them. That is what locks are supposed to do. Third, for all the inspiration I have taken from his writings, Marcuse's thought that there is a dominant logic or trajectory of technology is weakened by this analysis. Technological change has the potential to affect alienability, rivalry and exclusion cost in myriad ways. Xerox copiers, computers and the Internet have raised the exclusion cost for goods such as texts, audio recordings and images, at the same time that they have made them less rival. As a result, these items are less easy to control and less like commodity goods today than they were in Marcuse's lifetime. I paid good money my copy of *One Dimensional Man,* but readers of this article will have (very likely) accessed it for nothing on the Internet. Not surprisingly, those who benefited from the old material structure have moved quickly to encourage the enactment of formal legislation that would restore some of the rivalry and lower the costs they incur in excluding what they take to be unauthorized use.

Finally, even if technology should be questioned when alienability, rivalry and exclusion cost are affected, it is not at all obvious what the answer should be. Analysts who use the word "commodification" generally think that this kind of change is a bad thing, but economists who talk about reducing transaction costs generally think just the opposite. In both cases, there may be an understandable but false assumption that the material infrastructure of the world is relatively fixed, so that the processes in question always involve manipulations of law and policy. This assumption may then map transformations in alienability, rivalry and exclusion cost onto rather well-worn political ideologies. Hence, commodification is bad because it favors capitalist or bourgeois interests, while lowering transaction costs is always good because it allows rational agents to more successfully maximize the satisfaction of subjective preferences. Even if this is generally correct for changes in formal institutions, which I doubt, it will simply not do as a sweeping analysis of technical change. Lawrence Lessig's detailed studies of the way that technical codes affect alienability, rivalry and exclusion cost for software, the Internet and e-commerce suggest that when we question technology in this way, we will need to look closely at the actual implications of a specific technical change before we will be in a position to speak about whether it is good or bad (Lessig, 1999).

In conclusion, getting clear about alienability, rivalry and exclusion cost can help both innovators and philosophers of technology do some of things that they have long aspired to do better. In the case of technical innovation, these institutional factors represent parameters that go a long way toward predicting some of the most socially sensitive and historically contentious elements of a technical change. Be advised that such modifications will require careful planning and a well-crafted participatory process of design and implementation. For philosophers, they get us to at least some of the details that really matter when technical change occurs. A focus on alienability, rivalry and exclusion cost thus provides a promising way to integrate the philosophy, sociology and economics of technology, and to clarify some of the more obscure mechanisms that have been associated with technological determinism and social history. They also represent elements of specific technologies such as genetic engineering or information technology that serve as boundary objects linking alternative networks of actors, and bridging normative with classically technical domains. As such, they provide a focal point for the ethics of technology, and should be considered in any attempt to identify the elements of a novel technology that are most in need of deliberation and public discussion.

References

Commons, J.R. 1931. "Institutional Economics," *American Economic Review,* 21: 648-657.

Conway, G. 2000. "Genetically modified crops: risks and promise," *Conservation Ecology,* 4(1): 2. [online] URL: http://www.consecol.org/vol4/iss1/art2/

Feenberg, A. 1999. *Questioning Technology.* New York: Routledge Publishing.

Fink, D. 1998. *Cutting into the Meatpacking Line: Workers and Change in the Rural Midwest.* Chapel Hill: University of North Carolina Press.

Hickman, L. A. 2001. *Philosophical Tools for Technological Culture*. Bloomington: Indiana University Press.

Lessig, L. 1999. *Code: And Other Laws of Cyberspace.* New York: Basic Books.

MacPherson, C. B. 1962. *The Political Theory Of Possessive Individualism: Hobbes To Locke*. Oxford, Clarendon Press.

Marcuse, Herbert. 1966. *One Dimensional Man.* Boston: Beacon Press.

Misa, T., P. Brey and A. Feenberg, eds. 2003. *Technology and Modernity.* Cambridge, MA: The MIT Press.

Muir, W. 2004. "The Threats and Benefits of GM Fish," *EMBO Reports,* 5: 654-659.

North, D. C. 1990. *Institutions, Institutional Change and Economic Performance.* New York: Cambridge University Press.

Pickering, A. 2005. "Decentering sociology: synthetic dyes and social theory," *Perspectives on Science*, 13: 352-405.

Polanyi, K. 1944 (reprinted 2001). *The Great Transformation: The Political and Economic Origins of Our Time.* Boston: Beacon Press.

Thompson, E.P. 1971. "The Moral Economy of the English Crowd in the Eighteenth Century," *Past and Present,* 50 (February): 76-136.

Thompson, P. B. 2004. "The Legacy of Positivism and the Role of Ethics in the Agricultural Sciences," in *Perspectives in World Food and Agriculture 2004.* in C. G. Scanes and J. A. Miranowski, eds. Ames, IA: Iowa State University Press, 335-351.

Winner, Langdon. 1986. *The Whale and the Reactor: The Search for Limits in a Technological Age.* Chicago: University of Chicago Press.

Bug Tracking Systems as Public Spheres[1]

Joseph Reagle, Jr.
Department of Culture and Communication
New York University
The Steinhardt School of Education

Abstract
Based upon literature that argues technology, and even simple classification systems, embody cultural values, I ask if software bug tracking systems are similarly value laden. I make use of discourse within and around Web browser software development to identify specific discursive values, adopted from Ferree et al.'s "normative criteria for the public sphere," and conclude by arguing that such systems mediate community concerns and are subject to contested interpretations by their users.

1. Introduction

> "Last time I filed a bug report with KDE I got some snotty reply from some programmer who said I was wrong ([even so] the bug got fixed in the next release and was listed in the changelog)". - ErichTheWebGuy
>
> "I've submitted a number of bug reports and comments on existing bugs, and not only were they fixed promptly, but my privileges were raised so that I could close bug reports/mark duplicates/etc." - Anonymous Coward

The two comments above (Michael 2004) represent opposing positions within a discussion about reporting software bugs. A "bug tracking" tool permits one to identify, discuss, prioritize, close, and remove duplicate reports of system deficiencies. When most people think of bug and issue tracking software, if they do at all, they would probably think that it is a peripheral and mundane technology. Yet there are complex technical and social processes involved in addressing software bugs (Bork 2003).

As indicated by the frustration of "ErichTheWebGuy" above, a source of disagreement and even exit within open source communities is the handling of software bugs. In the open standards and software communities that this paper considers, the ways in which issues are represented with respect to their standing of consensus or dissension is affected by the processes, culture, and media of discourse (e.g. IRC, e-mail, Wiki, etc.).

Consequently, this paper is an exploration of how issue and bug tracking tools "embed," "embody," or "inscribe" cultural values of how a community should come to agreement, or even productively disagree. For example, what categories are available to describe the closure of a contentious issue? Or, how are the resource costs of reporting versus resolving a bug balanced?

2. Background: Values, Bugs, and Discourse

[1] Acknowledgment: I would like to thank Helen Nissenbaum for her comments on and discussion of this paper, and Nora Schaddelee for reviewing an earlier draft.

2.1 Values Embodied in Technology

In her provocatively entitled paper *Do Categories Have Politics?,* Lucy Suchman (1998) attacks the theory behind a Computer Supported Cooperative Work (CSCW) tool known as the "Coordinator." The operation of this tool was predicated on a foundation known as "speech actor theory." Suchman faults Winograd and Flores, the proponents of this theory and the designers of this tool because "the adoption of speech act theory as a foundation for system design, with its emphasis on the encoding of speakers' intentions into explicit categories, carries with it an agenda of discipline and control over organization members' actions" (Suchman 1998).

Terry Winograd responds to Suchman's question of "Why do computer scientists go about making up all these typologies of interaction?" (Suchman 1998) with this pragmatic reply: "The answer is relatively simple -- computer programs that we know how to construct only work with fully-rationalized typologies (be they bits and bytes or knowledge bases)" (Winograd 1998:109). Winograd acknowledges the potential problems of this process and notes: "The essence of using a tool well is knowing where, when, and how to apply it" (p. 111). This is reminiscent of the argument that guns don't kill people, people do. And while this essay will not address this complex question of a designer's responsibility – regardless of their intent – to all potential applications of their artifact, Winograd (p. 111-112) does offer some qualifications with respect to this type of design:

1. Explicit representation of intentions and commitments is more appropriate in some social/organizational situations than others.
2. The generation of representations can only be done successfully with the participation of the people who live the situations being represented.
3. It is a dangerous form of blindness to believe that any representation captures what is meaningful to people in a situation.

Yet, each one of these caveats merits a substantive discussion as well. Unfortunately, at this point I must avoid the particulars of that discussion to focus on what I hope the reader will accept as a presumption: that -- as Langdon Winner (1986) might say – "artifacts have politics." What are politics? Winner defines them as "arrangements of power and authority in human associations as well as the activities that take place within those arrangements" (p. 30). Not all politics are equally political, or political in the same way. In any case, my analysis is predicated on the simple point: technology is created and used by humans, and in both of those acts the technology interacts with and mediates the human/social sphere.

By way of example, Winner notes the wide Parisian thoroughfares that were intended to mitigate revolutionary barricades, the American university campuses built to facilitate easy troop movement and sniper positions during students protests, and the deployment of less efficient machines to displace unionized labor. Pinch and Bijker (1992) use the development of the bicycle as a case study for their argument for the social construction of technology. Latour (1992) argues that seatbelts and the "Berliner key", which requires one to close and lock the door behind oneself in order to retrieve the key, are delegations of human function and interest to an artifact. Weber (1985) describes the policy process whereby the U.S. Airforce altered height requirements in order to accommodate female pilots, who previously had been thought to be unsuited to the task. And Friedman and Nissenbaum (1997) identified numerous cases of "bias" in computer systems.

Clearly, technology design is an appropriate subject for policy analysis. Particularly for artifacts like automobiles, nuclear reactors, and bridges. But what of a filing system? Does a schema for

categorizing things deserve scrutiny? In *Sorting Things Out: Classification And Its Consequences*, Bowker and Star (1999:33) argue they do: "Systems of classification (and of standardization) form a juncture of social organization, moral order, and layers of technical integration." Bowker and Star described how a nursing intervention system was altered to recognize that the time spent with patients was an important activity, rather than an inefficiency:

> Information, in Bateson's famous definition, is about differences that make a difference. Designers of classification schemes constantly have to decide what really does make a difference; along the way they develop an economy of knowledge that articulates clearance and erasure and ensures that all and only relevant features of the object (a disease, a body, a nursing intervention) being classified are remembered. (Bowker and Star 1999:281)

Or, as Reagle (1999) noted in *Eskimo Snow and Scottish Rain: Legal Considerations of Schema Design*,

> In Judeo-Christian theology the first power granted by its God to man was the power to name, "Out of the ground the Lord God formed every beast of the field and every bird of the air, and brought them to Adam to see what he would call them. And whatever Adam called each living creature, that was its name." (Gen2:19-20). Designing a schema that others will use is -- in some sense -- an exercise of the power to name.

The example of Pluto being deprecated from the category of planet is a recent example of how contentious categorization can be. Designing the categories by which we interact with each other and our systems is bound to privilege some point of view, while muting others. Yet, not every system is a tool of sinister hegemonic forces with social implications far outstripping its technical scope. Sometimes the technology is very simple, as is its interface to the social world. How then, might one come to understand bug and issue tracking systems?

2.2 Bug Tracking

A bug tracking system is simply an issue tracking system about software bugs. Subsequently, I will use the term "bug tracking" generically unless there is cause to make a distinction. The reason I opt for "bug" over the more generic "issue" is because bug tracking systems are prominent in public use and as objects of discussion, and in practice many bug tracking systems track more than software bugs: they might also include proposals for new features (i.e. a wishlist).

One of the most well known bug tracking systems is Bugzilla. It is an open source project used to track bugs of other open source projects, most notably the Mozilla Web browser, a descendant of the Netscape browser. Open source projects produce software that is available in source code form and amendable to modification by others. Typically, the work process is open as well, so one can follow the discourse of the community in their e-mail, chat, or bug tracking conversations. Bugzilla (Mozilla 2002) describes itself as follows:

> Bugzilla is a database for bugs. It lets people report bugs and assigns these bugs to the appropriate developers. Developers can use Bugzilla to keep a to-do list as well as to prioritize, schedule and track dependencies.... Enter the tasks you're planning to work on as enhancement requests and Bugzilla will help you track them and allow others to see what you plan to work on. If people can see your flight plan, they can avoid duplicating your work and can possibly help out or offer feedback.

The Bugzilla system is a tool for collaboration, if for no other reason than to help avoid duplicate work. Shukla and Redmiles (1996) provide a succinct summary of the bug tracking process as a collaborative learning process and identify the stakeholders, including end-users, designers, implementers, and management. Additionally, sometimes a "bug-czar" or "quality assurance" person facilitates the processing of a bug through its "life cycle." Finally, while anyone could theoretically fix a bug, there is often a small group of individuals responsible for portions of code. A fix, or "patch," often comes from the core group since they know the code the best, or have the authority to mediate access to that code in the community's software versioning system.

In Bugzilla, when bugs are first submitted they are categorized as UNCONFIRMED, "this means that a QA (Quality Assurance) person needs to look at it and confirm it exists before it gets turned into a NEW bug" (Bugzilla 2004). When a bug is fixed it is marked as RESOLVED and given a resolution specified in (Bugzilla 2004):

> FIXED: A fix for this bug has been checked into the tree and tested by the person marking it FIXED.
>
> INVALID: The problem described is not a bug, or not a bug in Mozilla.
>
> WONTFIX: The problem described is a bug which will never be fixed, or a problem report which is a "feature", not a bug.
>
> LATER and REMIND: These are both deprecated. Please do not use them.
>
> DUPLICATE: The problem is a duplicate of an existing bug. Marking a bug duplicate requires the bug number of the duplicating bug and will add a comment with the bug number into the description field of the bug it is a duplicate of.
>
> WORKSFORME: All attempts at reproducing this bug in the current build were futile. If more information appears later, please re-open the bug, for now, file it.
>
> MOVED: The bug was specific to a particular Mozilla-based distribution and didn't affect mozilla.org code. The bug was moved to the bug database of the distributor of the affected Mozilla derivative.

When a QA person has confirmed the processing of a bug, the bug is marked as VERIFIED. When the software is "shipped" (the corrected version is available to end users) it is marked CLOSED though it may be REOPENED at a later time. As is evidenced by the number of categories and the deprecation of LATER and REMIND resolutions, this typology and process of tracking the bugs has evolved according to the experiences of the users of the system. Most bug tracking systems work in a similar way though there will be differences in their typology and processes.

While I am not able to provide a historical treatment of how the specific Bugzilla categories and processes came to be as they are shown above, I will identify some of the tensions that have prompted the development of such categories more generally and how those tensions are the subject of specific debates today. But to do this, I first want to briefly consider the types of values that might be embedded in the design of a bug tracking system.

2.3 Discursive Values in a Public Sphere

Bug tracking tools mediate a conversation between the user and the developer; the developer is responsible for addressing the item raised by the user. These designations are *roles*, for any person might easily be both a user and developer of a piece of software. (In fact, developers

frequently file reports against their own code.) These conversations are civil for the most part; for, unlike other scenarios such as a zero-sum trade dispute, both parties have substantive interests in common. It is in the user's interest to not encounter bugs; this is also in the developers' interest with respect to his own satisfaction and as a fellow user.

However, there can be differences between the roles. There may be particular bugs or features that a user wants fixed that is not a priority to the developer – she has her own interests and there is only so much time in a day. Or, when pressed for time or feeling confused about who is responsible for a particular bug, a user might submit a less than complete bug report.

Jürgen Habermas has influenced understandings of civic discourse with the concept of the *public sphere*, "a domain of our social life in which such a thing as public opinion can be formed" (Habermas 1991:398). While this framework seems rather disproportionate to small-scale discussions about software bugs, such a theory can provide characteristics of (sometimes contentious) discourse that are relevant to the questions I'm asking. For example, in *Normative Criteria for the Public Sphere,* Ferree et al. (2002) describe four forms of discursive tradition:

- representative liberal: elite dominance, expertise, and proportionality; a free marketplace of ideas with transparency, detachment, civility; with an outcome focused on closure (p. 206)
- participatory liberal: popular inclusion; empowerment with a range of communicated styles; avoidance of a premature closure (p. 210)
- discursive: popular inclusion; empowerment with a focus on dialog, mutual respect, civility (though impassioned) and merit based decisions; closure is contingent on consensus (p. 215)
- constructionist: privileges the periphery and oppressed; with a communicative narrative of empowerment; avoidance of premature closure (p. 222).

In some ways, this typology is inappropriate for the sort of technical conversations that are the subject of this paper because the voluntary character of open source development permits a different sort of relationship between discussion and action. In civic discourse, public opinion relates to governmental action via one of the forms above. In free/open source discourse, developers can and do argue that they need only satisfy themselves, those who disagree can do it their own way as well. (If it is a complement to what another developer has already done, it can be added; if it is an alternative, it will vie for adoption as a competitive "fork"). Yet this is a value itself – one sympathetic to the voluntary nature of much of the development. In cases when the community does want to condense a collection of opinion into a single policy many of the variables above, such as elite dominance, expertise, and transparency, are relevant to the analysis. In any case, the elements of each form are relevant, even if, for example, it is difficult to identify a perfect example of the constructionist form in bug tracking discourse. In the next section I present some real world cases in which these values are reflected and discussed in the context of bug tracking systems.

3. Method

This analysis is based on participation in the Web development community. Of most relevance to this paper, I was a user and bug reporter of various open source Web browsers; specifically, I followed the development of KDE's Konqueror browser (and desktop) and Apple's Safari browser, which was built upon Konqueror's open source HTML rendering engine. Ethnographic and archival data for this paper spans, roughly, three years (2000-2003) during which there was

much discussion of bug tracking strategies and the implementation of new tools. Sources include discussions from bug tracking systems, developer mailing lists and blogs, and a KDE news portal and discussion site. I did not attempt to interview participants, but instead, simply acted as one, while also making notes of my experiences: "A culture is expressed (or constituted) only by the actions and words of its members and must be interpreted by, not given to, a field worker" (Van Maanen 1988). All cited discourse is public and can be easily accessed on the Web.

4. Findings: Values, Strategies, and Voting

4.1 Values of Software Development and Bug Tracking

The very openness and explicitness of these Web browser bug tracking systems demonstrates a valuation of the principal of transparency. However, one must be careful in inferring intention on the part of designers towards a particular value. Langdon Winner (1986:29) argues that technologies like nuclear power are "inherently" political as they depend on particular types of political relationships. While this is a valuable insight, I am concerned more generally with the "social" and would avoid the essentialist characterization of "inherent." Instead, in many information designs, some technical values might be "sympathetic" to particular social values. For example, Lawrence Lessig (1999) discusses the technical benefits of the end-to-end architecture of the Internet, as well as the civil consequences this architecture had in facilitating free expression. Some might then infer that the designers of the Internet or Web started their projects with emancipatory purpose. Perhaps, but it might also be that this was an unintended consequence, a serendipitous consequence, or something which was not considered at all. (Such emancipatory inferences about intention are often made with the benefit of hindsight.) This is what then leads Lessig to argue that if we wish to preserve the values of the original Internet (both the open architecture and freedom of expression) we can no longer rely solely upon this sympathetic relationship because both the technology and social norm can come under attack; one should persist in open technical designs, and support freedom enhancing laws and social norms.

A critical and difficult job in the open software world is to compile the source code into easily installable *packages* that are then available as a *distribution* to the end users. This job is difficult for a number of reasons. The first of which is in managing dependencies. A benefit of open source development is that applications can share modular software functionality; yet, the ways in which these applications depend upon each other across multiple versions can be complicated. For example, a windowing desktop might depend on version 1.0 of graphical library to render the icons, but the latest version of a popular puzzle game might require version 2.0 of that same library! These two applications cannot easily coexist. When such problems occur the user is most likely to vent their frustrations upon the package maintainers, which is further complicated in that they may be the inappropriate recipient of the bug: it might be a problem with the package, but it also might be a bug in the original the source code.

The difficulties of this job are apparent in the Debian KDE desktop packaging community. (KDE is a windowing environment; Debian is a Linux distribution of easy to install packages or "debs.") In response to challenges about how the dependencies of a package were being handled, the package maintainer, Ivan Moore (2000a), responded "I'm really getting tired of this… I had to cut down on the number of bug reports I was getting and verify that the packages worked or didn't work." Eventually, Moore declared that he would stop maintaining the packages; Erik Severinghaus (2000) posted Moore's announcement to a KDE community Web site and editorialized:

> This happened with freshmeat.net a while ago, it has happened to countless projects,

> and I'm *tired* of dumbasses flaming developers/packagers/webmasters/whatever who volunteer their time to OpenSource projects. Stop bitching and fix it.

The next day Moore (2000b) relented:

> just a note. I have gotten a ton of email from alot of people who are upset about this. So far none of it negative. I want to make it clear that the negative comments come from about 1% of the community...it's just that this 1% is always the percentage that is the loudest. This only because they are saying that what you are doing is bad or wrong or [insert negative comment here]... Anyways...because of all the nice comments I had decided to make the KDE 2.0 potato debs available...or rather continue to make them available.

Yet, a similar event caused Moore to finally resign in January of 2002. The following year, Chris Cheney, one of Ivan's successors, was challenged for his performance and (presumed) inexperience. Charles de Miramon (2003) responded to the complaint as follows:

> I resent your ageism intruding into this. Chris is an excellent maintainer. Just because he doesn't have the time to answer the same question repeatedly to people who would rather complain than either fix the problem, or accept that they've done far, far less for the Debian KDE community than Chris, doesn't make him a bad maintainer… If you're so fanatical about this, go do something. Make a website. Talk to debian-desktop. Create a metapackage, whatever. It's more productive than the email you just sent.

From these threads we can clearly identify the values from Ferree et al. of resource efficiency (minimizing expended time), expertise (the ability/merit of the maintainer and user), proportionality (the effect the 1% minority might have on morale), self-reliance/commitment (exhorting others to contribute), and mutual respect (providing positive feedback when needed).

4.2 Wizards and Strategies

In September 2002, the KDE bug system was switched over to a Bugzilla implementation with a KDE specific five-page bug reporting wizard. Prior to the use of Bugzilla and wizard, bugs were submitted via a single complex form. In an effort to encourage complete and unique bug reports, the wizard requires the completion of information such as a version and distribution, and presents the user with a set of existing bug reports that may be relevant. However, some frequent users considered the five-page wizard to be tedious. (The danger is that if a system is difficult to use, it can yield fewer legitimate reports.) Sebastian Laout (2004) submitted a bug report against the wizard itself: "Posting a bug in bugs.kde.org is a pain" and included a step-by-step analysis of the inefficiencies of the wizard process. However, presently, the bug's status is RESOLVED with a resolution of WONTFIX. Daniel Naber responded, "We *need* the wizard so that people stick at least to *some* rules. Otherwise we will drown in duplicates and reports that are even worse than now. If you have a better idea for the wizard, send patches." This is again demonstrative of the values of efficiency and self-reliance/commitment.

However, even within a perfectly efficient bug reporting system, the tension of differing priorities would remain. Dave Hyatt (2003), a lead developer of the Safari Web-browser for the Macintosh, noted an amusing strategy of bug submitters vying for developer attention:

> I love the tactics some people use when filing bugs. In particular the tactic of saying something inflammatory in order to goad the receiver of the bug into fixing it. You see this a lot in Bugzilla, and also in reported Safari bugs.

> Here are some of my favorite phrases (for your enjoyment). Let X = the browser of your choice. Let Y = any other browser.
>
> (1) The Promise - "The lack of this feature is the one thing that keeps me from switching to X."
>
> (2) "I can't work under these conditions. I'll be in my trailer." - "I can't believe you broke this! That's it! I'm going back to Y!"
>
> (3) Playing the EOMB Card - "How can this be broken? Every other modern browser gets this right."
>
> (4) Impatience - "Months have passed, and this bug still hasn't been fixed! What's the holdup?"
>
> (5) Overeagerness - "Still broken." (2 days later.) "Still broken." (2 days later.) "Feature still doesn't work. (2 days later.) "Broken in build from mm/dd/yy."
>
> The Safari team has actually started using the term EOMB as a way of referring to all other modern browsers. ;)

In order to give a voice to the user community, and limit minorities from using morale damaging strategies, some free/open software communities have implemented bug voting schemes.

4.3 Voting and "Democracy"

In a typical bug voting scheme, each registered reporter is allocated a fixed amount of points that they can spend on bugs, up to some ceiling per bug or application. The front page of the KDE bug reporting system includes reports such as weekly summary statistics, the most hated bugs, the most wanted features, the most frequently reported bugs, report counts by ownership, severity, and priority. (An additional feature of Bugzilla is that an UNCONFIRMED bug with a sufficient number of votes can be automatically elevated to NEW without the intervention of a quality assurance person.) This model is reminiscent of Ferree et al.'s "representative liberal" form wherein the media serves the purpose of ensuring the accountability of the representatives via transparency. Yet, different communities interpret the meaning of votes differently. Or, as Brey (1997) argues technical systems are subject to "different interpretations, not only of its functional and social-cultural properties but also of its technical content, that is, the way it works" (Brey 1997).

The Mozilla community quality advocate, Asa Dotzler (2002), has stated, "Votes aren't ignored but at the same time they're not the deciding factor in what gets fixed." He noted that votes are disproportionately spent on feature requests, disadvantaging critical bug reports; that those who file bug reports are a tiny fraction of all Mozilla users; and he argues bug reporters are probably not representative of the larger community. Furthermore, the voting scheme is simplistic (e.g., users can't vote against a feature).

Another common point of discussion is whether one should solicit others to vote on a particular bug. Aaron Seigo (2003) objected to this practice:

> if i may suggest, the best way to make voting on bugs.kde.org absolutely worthless is to recruit people wholesale to vote for various random bugs by posting them off-topic to places such as theDot. such campaigning distorts the statistical relevance inherent in the process. while you may achieve a surge in votes for your pet bug, you'll be doing a disservice to all the other bugs that have garnered votes "the hard way" even though

> those votes are probably much more relevant/important

Such a viewpoint represents a pluralistic view of a public sphere: each user should represent her own position, and the role of compromise and representation is seen as ultimately corrupting. Some participants note that these discussions begin to take on the character of "real-world" politics:

> "IMHO this is getting as annoying as the campaigning of political parties" (Loose 2003).
>
> The comparison to elections and advertising is truly astonishing given that campaign reform and an attempt to end undue influence returning to a "one person, one vote" ideal has been at the forefront of politics for years (Laffoon 2003).

An interesting issue that arises when one attempts to assess, for one's own satisfaction, which position on voting is "correct" is that there is no right or wrong; instead, what can be important for the cohesion of the community is the degree to which one of those interpretations is commonly held.

5. Conclusion

Bug tracking systems are, at first glance, seemingly boring and of little relevance on questions of community and discourse. On second glance, they might be seen as a media through which the community discusses and prioritizes issues important to it, but only in a narrowly technical way. In this paper I show that bug tracking systems mediate tensions between members of a software community. Adopting Ferree et al.'s "normative criteria for the public sphere" I identify within the KDE community the importance of the values of resource efficiency, expertise, proportionality, self-reliance/commitment, and mutual respect. When the KDE community became aware of the tensions between stakeholders and such values (e.g., users attempting to receive attention and developers responding "do it yourself") they deployed mechanisms such as bug voting. However, this prompted discussion on the appropriateness of campaigning and vote trading! From this, I conclude that this case exceeds the theoretical framework of "embedded," "embodied" (Grint and Woolgar 1995) or "inscribed" (Latour 1992) values. Instead, this case highlights the importance of ongoing interpretation (Pinch and Bijker 1992) in understanding the *meaning* of technology -- going beyond designers' intention.

References

Brey, P. 1997. “Philosophy of technology meets social constructivism,” *Techne: Journal for the Society for Philosophy and Technology:* 2(3-4).

Bork, J. 2003. “Anatomy of a bug,” Incessant Ramblings, http://headblender.com/joe/blog/archives/microsoft/001280.html

Bowker, G. C., and S.L Star. 1999. *Sorting things out: classification and its consequences*. Cambridge, MA: MIT Press. http://www.istl.org/00-winter/review2.html

Dotzler, A. 2002. “Mozilla 1.2.1 Coming Soon,”mozillaZine, http://www.mozillazine.org/talkback.html?article=2702

Ferree, M. M., W.A. Gamson, J. Gerhards, D. Rucht. 2002. “Normative criteria for the public sphere.” In *Shaping Abortion Discourse*. Cambridge: Cambridge University Press.

Friedman, B. and Nissenbaum, H. (1997). “Bias in computer systems,” In B. Friedman, ed., *Human Values and the Design of Computer Technology,* New York: Cambridge University Press, 21-40.

Grint, K., and S. Woolgar. 1995. "On some failures of nerve in constructivist and feminist analyses of technology," *Science, Technology, and Human Values,* 20: 286-310.

Hyatt, D. 2003. "Bug Guilt Trips," Surfin' Safari, http://weblogs.mozillazine.org/hyatt/archives/2003_11.html#004358

KDE News. 2002. "KDE Switches To Bugzilla,"http://dot.kde.org/1032319933/

Laffoon, E. 2003. "Re: OT: pls vote for this bug,"KDE News http://dot.kde.org/1058371499/1058390231/1058392962/1058425970/1058461094/

Latour, B. 1992. "Where are the missing masses? The sociology of a few mundane artifacts," In W. Bijker and J. Law, eds., *Shaping Technology/Building Society,* Cambridge, MA: MIT Press.

Lessig, Lawrence. 2000. *Code and Other Laws of Cyberspace,* New York: Basic Books.

Loose, C. 2003. "Re: OT: pls vote for this bug," http://dot.kde.org/1058371499/1058390231/1058392962/1058425970/1058429014/

Malone, T. 1998. "Commentary on uchman article and Winograd response." In B. Friedman, ed., *Human Values and Design of Computer Technology*, chapter 6, Oxford: Cambridge University Press.

Michael. 2004. "Announcing the KDE Quality Team Project," http://slashdot.org/article.pl?sid=04/03/02/1924204

Moore, I.E. 2000a. "Re: KDE & apt,"http://lists.debian.org/debian-devel/2000/debian-devel-200010/msg00445.html

Moore, I.E. 2000b. "Re: KDE Linux Packaging Project Taken Down," http://dot.kde.org/971680096/971755828/

Mozilla. 2002. "Bugzilla,"http://www.bugzilla.org/

Pinch, T., and W. Bijker. 1992. "The social construction of facts and artifacts: or how the sociology of science and the sociology of technology might benefit each other." In W. Bijker and J. Law, eds., *Shaping Technology/Building Society,* Cambridge, MA: MIT Press, 17-50.

Reagle, J. 1999. "Eskimo Snow and Scottish Rain: Legal Considerations of Schema Design. Reference: W3C Note 10-September-1999," http://www.w3.org/TR/1999/NOTE-md-policy-design-19990910.html

Seigo, A. 2003. "Re: OT: pls vote for this bug," http://dot.kde.org/1058371499/1058390231/1058392962/

Severinghaus, E. 2000. "Re: OT: pls vote for this bug," http://dot.kde.org/1058371499/1058390231/1058392962/

Shukla, S., and D. Redmiles. 1996. "Collaborative learning in a bug-tracking scenario," In *Conference on Computer Supported Cooperative Work* (CSCW 96), Association for Computing Machinery.

Suchman, L. 1998. "Do categories have politics? The language/action perspective reconsidered," In B. Friedman, ed., *Human Values and Design of Computer Technology*, chapter 4, Oxford: Cambridge University Press.

van Maanen, J. 1988. *Tales of the field: on writing ethnography,* Chicago: University Of Chicago Press.

Weber, R. 1985. *Manufacturing gender in the cockpit design,* Open University Press.

Winner, L. (1986). "Do artifacts have politics?" In *The Whale and the Reactor*, pages 18-39. The University of Chicago Press, Chicago.

Winograd, T. 1998. "Categories, disciplines, and social coordination," In B. Friedman, ed., *Human Values and Design of Computer Technology*, chapter 5, Oxford: Cambridge University Press.

Virtual Models and Simulations: A Different Kind of Science?

Peter R. Krebs
School of History & Philosophy
University of New South Wales

Abstract
The personal computer has become the primary research tool in many scientific and engineering disciplines. The role of the computer has been extended to be an experimental and modelling tool both for convenience and sometimes necessity. In this paper some of the relationships between *real* models and *virtual* models, i.e. models that exist only as programs and data structures, are explored. It is argued that the shift from experimenting with real objects to experimentation with computer models and simulations may also require a new approach for evaluating scientific theories derived from these models. Accepting the additional sets of assumptions that are associated with computer models and simulations requires 'leaps of faith', which we may not want to make in order to preserve scientific rigor. There are problems in providing acceptable arguments and explanations as to why a particular computer model or simulation should be judged scientifically sound, plausible, or even probable. These problems not only emerge from models that are particularly complex, but also in models that suffer from being too simplistic.

Introduction

In a recent volume by De Chadarevian and Hopwood (2004) a number of authors present some of their views on 3-dimensional models and the role such models play in science from a historiographical perspective. The various models discussed have in common that they are mostly *material* things, i.e. models made of clay, wood, plasticine and the like. In the last few decades computers have revolutionized scientific modelling, and the notion of *model* has changed. The use of 'computer models' does not just add another kind of model to the array of 'traditional' artifacts. In some disciplines computers have become *the* modelling tool, rather than merely playing a supplementary role. Indeed, in the field of Cognitive Science the computer model *is* the 'traditional' model, given the underlying computational theory of mind. Not only are many of the characteristics of computer models and simulations entirely different from material models, but the way we interact with models changes as a consequence.

Now, there are no longer real objects to probe, to measure or to collect, and all of our activities target mere *representations* of the world, i.e. mathematical abstractions, and computations with these representations (symbols). Moreover, a new layer of '*virtual reality*' is often created with the aid of various visualization techniques. Experimentation with such models in an interactive and 'interfering' way that Hacking (1983) and Harré (1970) ask for is not possible. Instead, the experiments are conducted in the domain of the virtual and the computational paradigm. Yet computer models are sometimes deemed to be real world objects in the same way the objects that are modelled are real world objects.

Not long ago, the concept of simulation "invariably implied deceit" (Keller, 2003). I think that this sentiment also applied to the term model, albeit to a lesser degree. Simulations and models were thought of as merely mimicking, or faking, the real world. While modelling has become a widely used technique in almost any imaginable discipline, the term is still often associated with a certain amount of incredulity, or, skepticism. For every model that shows *A*, there seems to be

always an alternative model showing *B*, and it is significant that we quite often hear the expression 'it is only a model'. In contrast, the term *model* is also used to denote standards and even perfection: the *model* husband (Jordanova, 2004). Expressions like 'virtual models' and 'virtual experiments' might be preferable, because the term *virtual*, and in particular *virtual reality*, seem more positive and are usually associated with cutting edge computing and Artificial Intelligence (AI).

The arrival of modern computing machinery in the 1950s and the proliferation of inexpensive and very powerful PCs since have led to a revolution in terms of what kinds of models and simulations can be implemented. Computer models and simulations make use of many advanced techniques that introduce new, often exciting, ways to present aspects of models to scientists, science communicators and 'consumers' of science alike. Computer generated images (CGI) not only changed the way we think about pictures and movies, but also about how theories are formed. New and innovative methods have been devised to present data in both scientific and non-scientific contexts. There is a variety of powerful methods for visualization and presentation available, and many applications of these techniques have made their way into textbooks and journals in the form of illustrations and graphical representations. With advanced image processing techniques, it is not only possible to alter and to enhance pictures, but it is also possible to render images of phenomena that are not visible, or may not exist at all. Many of the computer aided experiments and visualizations may be helpful in understanding complex phenomena because "visualizations contribute to 'amplify cognition'" (Araya, 2003). However, due to some reservations about the validity of computer simulations as experiments and methods to gain 'scientific knowledge', it seems that virtual models introduce a different set of issues concerning scientific rigor. Accepting virtual models and virtual simulations as experimental or empirical tools in science, will force us to adopt some new form of '*virtual* scientific method'.

Building Models

During the process of building a computer model or simulation, several transformations, or translations, take place. In the first instance there is a transformation of the (sometimes) observable phenomena or theoretical entities and the relationships between them into their corresponding mathematical entities. The result is a mathematical model that has been described as an intellectual construct, or, a mathematical object (Jorion, 1999). Then there is a translation of the mathematical structures into computational entities that are designed to deal with the complexities of the calculations in an appropriate, effective and efficient manner. The third transformation takes place when the data, which has been generated or transformed by models, is translated into a format that is more easily interpretable by the experimenter. In models and simulations, where large amounts of data are involved, additional steps are usually taken to present the data in some sort of visual form. The final transformation occurs when the model is reinterpreted in the language of the initial problem, question or theory.

Mathematical Models

A mathematical model is an intellectual construct that is based on a mathematical object, which "does not tell anything about the world" (Jorion, 1999). Jorion believes that mathematical objects, without sensible interpretation, are all about syntax, and any of their meaning derives entirely from its structure. The inherent meaning held by a mathematical object is that

> […] some of the symbols which constitute [the mathematical object] impose constraints on others, some have no more meaning than the set of constraints they are submitted to (Jorion, 1999, 2).

This view of a *symbol* is comparable with that of a representational system (RS) of 'Type 1' suggested by Dretske, who says about these kinds of RSs that they are "*doubly* conventional: *we* give them a job to do, and then *we* do it for them" (Dretske, 1988). However, a mathematical model becomes more than just a collection of meaningless symbols, provided that a sensible interpretation is possible. Jorion goes as far as to say that the mathematical model and the part of the world that is modelled are *isomorphic*, provided that the interpreted model is meaningful, i.e. the model "makes sense".

> The benefits of a mathematical model for world comprehension are the following: if an interpreted mathematical model makes sense, then it is reasonable to assume that the type of relations which hold between the symbols in the model hold also between the bits of the real world which are represented in the interpretation of the model (Jorion, 1999, 3).

The analysis of many models, in terms of initial assumptions and the claims made later, reveals that there are many different opinions on what "makes sense". Artificial neurons, for example, have very little in common with real neurons: they differ in their external functionality, their behavior, and their architecture. Other than a gross similarity in that they transform (integrate) several inputs to one output, they really share only the name. The isomorphism of biological neurons and mathematical neurons can barely be described as an 'approxi-morphism', let alone as an 'isomorphism', but the question of whether it "makes sense" to employ simplistic artificial neurons in cognitive models or not, is certainly not asked often enough. Psillos (1999) refers to 'modelling assumptions' that reflect the relationship between the model and the target physical system. He thinks that

> [f]ar from being arbitrary, the choice of modelling assumptions for [the target system] *X* is guided by *substantive similarities* between the target system *X* and some other physical system *Y*. It is in the light of these similarities that *Y* is chosen to give rise to a model *M* of *X* (Psillos, 1999, 140).

I believe that these "substantive similarities" also capture the nature of the relationships between the mathematical description (model) and a theoretical entity, provided certain conditions are met. Some of these conditions are discussed later.

How do we derive a mathematical description of some relationship among physical (or mental) entities? There are several conceptual transformations and processes involved. In the following paragraphs I discuss some of the issues concerning *abstraction*, *formalization*, *generalization*, and *simplification*, because these operations should be considered fundamental steps in the process of building (or constructing) mathematical models.

Abstraction

Mathematical models refine the real world by introducing an element of abstraction. It is clear that a model should be simpler than that which is to be modelled. The process of abstraction involves several practices, all of which widen the gap between the sometimes observable

phenomena and an idealized description in mathematical terminology. Stufflebeam (1998) suggested that his cat Sophie's behavior, when dropped from two feet, "satisfies the distance function $D(t)=½gt^2$ ". The abstraction here includes the reduction of the cat Sophie to a point mass in Newtonian physics. The observable behavior of the cat Sophie in free fall differs from the idealized point mass. In fact, Stuffelbeam's description of Sophie's behavior as $D(t)=½gt^2$ does not involve Sophie at all. Abstraction is the process of defining a general and idealized case for relationships between entities and processes. In the distance formula, *g* stands for the acceleration, and if we substitute the values for *g* of 9.8 m/s^2 or 32 ft/s^2 then we get a reasonable approximation of the conditions on Earth. However, we can also find the appropriate values for this model to work on the moon or on Mars. The distance formula holds *anywhere* for *any* object, provided we have the correct value for *g*. The most important aspect here is the introduction of placeholders like $D(t)$, which is the abstract notion of 'the distance of something at a particular moment in time'. This placeholder, or symbol, can now be manipulated within a formal system, like mathematics in this case. The introduction of symbols may put constraints on the type of operations and the methods for the model. For example, a sigmoid squashing function is selected in neuron models (perceptrons), because (1) the function's behavior is close to that of the step function at some level and (2) the function is differentiable at every point. While the qualities of the step function are desirable for the implementation of a neuron's functionality, some mathematical procedures (the back propagation of error algorithms for learning in this case) *require* that this function is differentiable everywhere. The point is that the mathematical methods that make up the model, or play an essential part in the model, are likely to dictate the kind of mathematical structures of the model at some level.

Formalization

The second and usually difficult process in building a mathematical model is that of *formalization*. A mathematical description of entities and the relationships that hold between them can only work as a useful model if there is a sufficient precision of terms. In areas of elementary Physics, like Newtonian dynamics, models work well because terms like *mass*, *velocity* and *force* and the interaction between these concepts are defined within a formal system, based on axioms. This is not the case in other scientific endeavors. The difficulty in Cognitive Science, for example, is that many terms describe mental things, like beliefs, behaviors and linguistic concepts, rather than physical things with properties that can be described and defined easily. Moreover, for mental concepts, we do not have clearly defined relations or processes to manipulate such concepts. Green (2001) suggests that some of the apparent success of connectionist models is due the lack of precision of terms (*vagueness*) and insufficient explanations of what it is that is actually modelled. The question is whether beliefs or behaviors can be modelled successfully, if it is not possible to provide a formal description of what we want to model. However, formal representations of a belief, for example, are needed in a computer program, because we need some way of encoding this concept. I suspect that formal descriptions of mental events, if it is at all possible to produce such descriptions, will not be in terms of simple placeholders. They will have to be either simple and relatively vague, or they will be very complex in order to provide some exactness and precision. But there is a catch: on the one hand there has to be sufficient precision to build a good model, on the other hand, precision in terminology and in detail makes it harder to build models that remain simple. Formalization ought to eliminate many of the 'soft' assumptions and descriptions about mental concepts. However, mental concepts are not easily defined or described in formal terms. For example, experience with the representation of knowledge in many applications in the field of AI have

shown, that it is very difficult to encode '*facts*' and the associated rules. In these situations we have to face the additional problem of also having to encode the subjective *degree* of belief and quite likely fuzzy representations about what *is* believed. It is clear that we cannot choose suitable sets of symbols, sets of rules of inference and transformation rules for mental concepts in the same way we can choose $D(t)$.

Generalization

For some models it is desired that they work well for a theory about something "in principle", rather than to target a particular instance of the theory in question. In other words, the model has to be able to produce data, if the model is designed to predict some cognitive behavior in humans, for *humans*, rather than the behavior of *Lucy* or *Bob*. There is, of course, the added problem of validating the model. In order to measure the success of the model, we need to compare data from the model with real data. The *real* data in this case has to be statistical data, because averaging data from many individuals can provide us only with generic *human* data. Generalization is usually achieved by omitting detail and allowing for a very broad interpretation of results. The danger here is to make models so general that they no longer capture the complexity of the theory or issue to be modelled (Krebs, 2005; Krebs, 2007). For example, the general formula for falling objects (e.g. cats) based on simple Newtonian physics is not a sufficiently precise model for what happens to a parachute jumper in free fall. For the latter much more specific case, it is important that drag and terminal velocity are considered in the model.

Simplification

One of the many criteria defining what makes a 'good' model is that the model is easier to work with. One way of making models easier is to simplify things, which can be achieved by disregarding details or external (environmental) issues that influence the model. For Sophie, the distance traveled by a free falling object on earth can be modelled using $D(t)=½gt^2+v_0+D_0$ where g is the acceleration of about 10 ms^2, and v_0 is the vertical velocity at the beginning of the time interval t. $D(t)$ gives us the distance after the time t from the position D_0, the position of the object at the beginning of the time interval. This is an 'easy to work with' model, because we do not take into account, among many other things, that (1) the acceleration is only *approximately* 10 ms^2, and (2) the atmosphere causes drag. Even when taking drag into consideration, the mathematical model of Sophie's behavior is still crude, because we have not considered the Reynolds numbers, the variation of the gravitational force over geographical regions, and many other perturbations. If we take drag into consideration we need to know that drag itself depends on, among other things, (1) the shape of the object and (2) air density. But, the density of the air is dependent on the temperature and the humidity, and the Reynolds numbers depend on the velocity of the object (cat), its shape and size, its surface, and so on. In the case of Sophie, the problem can not be fully described, because the cat could and would change its shape and therefore many parameters during the free fall.

At some point the model will become so complicated that it is no longer easy to work with, because the model is more difficult to understand than the original problem. $D(t)=½gt^2+v_0+D_0$ is likely to be sufficient as a model for most 'dropping cat problems'. I consider the task of simplification to determine what must be included in the model, and what kind of detail can be omitted, the most difficult challenge. As computers become more powerful, the computational

complexity of models can be increased, which in turn, as one would expect, will increase the quality and power of the models. But this is not necessarily the case. A very complex model is no longer easy to use, and an increased complexity can also be an indication that the model needed many additions to explain or produce acceptable predictions. In the same way Tycho Brahe's model of the universe needed more and more epicycles to 'keep up' with the data that was gleaned from actual observations. It might turn out that the model is just not good enough to explain things adequately.

Experiments

It has been suggested by some Philosophers of Science, e.g. Harré and Hacking, that experimentation is not merely about the observation of phenomena and subsequent inferences to the explanatory theories. Instead, experimentation is also about observing and *interfering* with the objects in question (Hacking, 1983). The ability to manipulate objects is an essential and integral part of the process of experimentation, which is "to create, produce, refine, and stabilize phenomena" (Hacking, 1983). The close connection between the experiment, a material model and the real world is also a key requirement in a definition of the term *experiment* offered by Harré, who says that

> [a]n experiment is the manipulation of [an] apparatus, which is an arrangement of material stuff integrated into the material world in a number of different ways (Harré, 2003, 19).

Harré suggests that the experimental setup (apparatus) is either an instrument that can tell us something about the world due to the causal relationships between the setup and the 'states of the world', or it is a "domesticated version of the systems in the world"(Harré, 2003, 26).

The kinds of experiments that fit the criteria suggested here are associated by some, naïvely perhaps, with what actually happens in a laboratory. These are the kinds of experiments that remind us of our high school days. However, it has become obvious that the vast majority of experiments are different from this stereotypic view (Morgan, 2003). When we conduct experiments with computational models and simulations, there are no materials that could possibly be manipulated. The material, the apparatus and the process of interference are all replaced by data structures and computational processes. The nature of the entities and the phenomena that are the points of interest in the field of Cognitive Science, for example, dictates that models and simulations are often the only way to do any experimentation at all. In Cognitive Science, the experiment is moved into the realm of the *virtual*, not just for convenience, but more often than not, out of necessity.

Virtual Experiments

Elements of computation can be part of a causal chain. Imagine an experimental setup where micro-electrodes are used to measure some voltage changes in a living cell in response to some stimulus introduced with another set of micro-electrodes. Instead of using a voltmeter that is built around a mechanism involving a coil, a magnet and a pointer with a dial, the voltage is displayed on a computer screen. The voltage differential at the electrodes is converted into a digital signal so that a particular voltage is represented (encoded) as a binary bit pattern. This data is fed through one of the computer's input/output channels, and a program performs the task to convert and display the data as a series of figures, i.e. numbers, on the screen. There are, in principle at least, no difficulties in explaining the causal chain between the number on the screen and

electrical potential at the micro electrodes. The numbers on the screen are elements of a Type II RS in the Dretskian theory. RSs of Type II are grounded in the real world in that their power to *indicate* is linked to causal events in the real world. Linking meaning to causal events has also been suggested by Russell, who defines "causal lines" as

> […] a temporal series of events so related that, given some of them, something can be inferred about others whatever may be happening elsewhere. A causal line may always be regarded as the persistence of something - a person, a table, a photon, or what not. Throughout a given casual line, there may be consistency of quality, consistency of structure, or gradual change in either, but no sudden change of any considerable magnitude. I should consider the process from speaker to listener in broadcasting one causal line: here the beginning and the end are similar in quality as well as structure, but the intermediate links - sound waves, electromagnetic waves, and physiological processes - have only a resemblance of structure to each other and to the initial and final terms of the series (Russell, 1948, 477).

The meaning (i.e. our interpretation of the semantic content) is not bound to the real world in the same way. The power to indicate something about the real world has to be recognized by the observer of the sign. Scientific instruments, thermometers, or voltmeters, indicate temperature, electric potential and similar properties and phenomena. They function by exploiting (detecting) a known physical phenomenon. Instruments provide the observer with a representation of the state or relation of that phenomenon through a series of often complex transformations. For example, it is a property of the real world that the volume of a quantity of metal varies with temperature. A suitable arrangement of levers and a pointer on a dial can be used to exploit the relationship between temperature and volume to create an instrument that will indicate the temperature with some accuracy. There is a distinction between what the instrument indicates and what the observer believes that indication *means*. The pointer on the dial will only be meaningful to someone who *knows* that this instrument is indeed a thermometer. The instrument will *indicate* the temperature quite independently from the observer. To be an indicator of some particular property of the real world, the causal relationships must be maintained and the observer must attach the right kind of interpretation in terms of the indicator's meaning. Dretske explains that

> [i]f a fuel gauge is broken (stuck, say, at "half full"), it *never* indicates anything about the gasoline in the tank. Even if the tank *is* half full, and even if the driver, unaware of the broken gauge, comes to believe (correctly, as it turns out) that the tank is half full, the reading is not a sign - does not mean or indicate - that the tank is half full. Broken clocks are *never* right, not even twice a day, if being right requires them to *indicate* the correct time of day (Dretske, 1988, 308).

In the suggested example, i.e the computer indicating voltages and the like, the computer is an integral component of the experimental setup, but the computer is not implementing a *virtual* model. Consider the following changes to the experiment. The computer program is now modified to read the pattern and to display the corresponding number every second, and as an additional feature, the program records the time and the values from micro-electrodes in the machine's memory. The information in the memory can also be *replayed* so that the sequence of numbers is displayed on the computer screen in one second intervals. Essentially the computer is now *simulating* the original experiment by re-playing what happened earlier. Is there now a problem in causally linking the patterns in memory to the micro electrodes? I suggest that there is not. There is only a time delay that has no bearing on the 'causal chain', because the data for

the simulation has been obtained by means that is (was) in principle 'causally' traceable. The experiment continues and a mathematically minded researcher recognizes that a pattern seems inherent in the data. She pushes the data through her favorite statistics program on her computer and finds a very good fit of the data for some function $f(x)$. Because of the difficulties when dealing with living neurons in these kinds of experiments, it is decided to build a model of the neuron's function based on $f(x)$. The neuron and all micro-electrodes are dispensed with, and the stimulus is now generated within the computer model varying the values for x in the domain of f, and the corresponding values $f(x)$ is displayed on the screen. Obviously, we can now easily produce many more data points to fill any gaps. The data that comes out of the model is different in terms of its origin and therefore also different in terms of what conclusions we can draw about f.

While the causal connections of some symbol *could* be traced, at least in principle, the *actual* connection to the real world is merely *assumed*. This assumption, because it is at least in principle traceable, maintains or promotes the symbol to an element of a Type II RS. The relationship between the real world and representations of the world is of great importance in the context of models and simulations. The value of a model as a representation of the real world and any insights into the working of the world by investigating properties of the model depends on the kind of representations the model employs. A *meaningless* representation, in the sense that it can represent arbitrarily *anything*, can and will render the entire model meaningless, unless there is a syntactically correct procedure (probably a causal chain) to tie these representations down. We can accept that some mathematical constructs and computer programs produce useful data (predictions) or that they perform suitably in the context of a particular problem, without having any similarities to the entities and relations of the problem at hand. However not all such 'models' may be able to offer any explanations or insights in another domain. For example, some computer programs, which may be designed to follow principles from the field of AI, perform the task of reading aloud some arbitrary text surprisingly well. However, these programs do not offer anything in terms of how a human being performs the same task - these programs are *faking* it, even if Artificial Neural Nets are involved (Krebs, 2005).

A model, or representational system, that is to function as a representation of the real world ought not to contain any Type I elements. In addition, representations of Type II, by definition, must not have gaps or uncertainties in the causal chain linking them to the real world. A thermometer is only a thermometer if it has the power to indicate the temperature. Some apparatus may well indicate the temperature provided certain other conditions are given. An example will illustrate this point. Imagine a partially inflated balloon that is connected to a pressure gauge. The volume of air and the air pressure inside the balloon will change with the ambient temperature and the ambient air pressure. This setup will function as a thermometer, *if* the ambient air pressure is kept constant. However, if the temperature is kept constant and the ambient pressure is allowed to vary, then the instrument will indicate pressure. This simple instrument has the power to indicate either *temperature* or *pressure*, that is, the setup can function as a thermometer *or* a barometer. A scientific, or a merely *usable*, instrument would have to be engineered so that the relationships between pressure, temperature and volume are exploited. But the power to indicate one or the other must be constrained through appropriate means to guarantee an indication of either only pressure or only temperature.

Type II representational systems contain *natural signs* that are objectively connected to the real world and their power to indicate something about that world is exploited by *using* their natural meaning (Dretske, 1988), because

> [i]n systems of Type II, natural signs take the place of symbols as the representational elements. A sign is given the job of doing what it (suitably deployed) can already do (Dretske, 1988, 310).

It is important to note that there is no intentionality associated with this type of representation. However, the potential intentionality (the meaningful interpretation) is constrained by the causal links to the real world. The variation in volume of metal, for example, may be due to the change of temperature, but this variation in volume cannot be reasonably attributed to the colour of the paper it is wrapped in.

Computer models and computer simulations have become tools for science in many ways. AI and Computational Neuroscience are special cases among the 'hard' sciences in that computation *is* the very nature of their activities. Other sciences might employ computational models and simulations as tools, however chemistry, for example, is essentially about elements, molecules, compounds, plastics or pharmaceuticals, even if computational models and simulations play a role in chemical research. AI, in contrast, takes computational models to simulate, even replicate, cognitive functions that are *computational* themselves. This would certainly be the case if the assumption that cognition *is* computation is true. If it turns out that cognition is merely computable, then AI would still be entirely about computation, but the contributions to Cognitive Science would need additional justification.

Three distinct types of models have been identified where, (1) computers are used to deal with theories and mathematical abstractions, which would otherwise be computationally intractable, (2) computers provide responses (data) in 'what-if' simulations, i.e. the behavior of a real world physical system is simulated according to some theory, and (3) computers simulate the behavior of non-existing entities, for example the simulation of artificial life (Keller, 2003).

The role of computational models and 'virtual experiments', i.e. simulations, as contributors in the framework of empirical science are of particular importance. This holds especially for Cognitive Science because many of the objects of inquiry in Cognitive Science cannot be observed directly or mediated by scientific instruments. Consequently, models and simulations are often the *only* method available to the scientist. It has been argued that computer simulations are essentially extensions of numerical methods, which have been part of scientific reasoning for a long time (Keller, 2003; Gooding, 2003). Human beings do not reliably maintain accuracy when they have to deal with a large quantity of numbers, and digital machines are much more efficient at doing logical and numerical calculations. The recognition of patterns and structures is much more the domain of human beings. The work of analysis and interpretation of patterns, whether these are observed directly or whether these are produced by a machine, remains largely the task of the scientist. Ziman (2000) suggests that what can be known to science is restricted to what is known to scientists, when he says that "[a]n empirical scientific fact originates in an observation - an act of human *perception*"(Ziman, 2000, 102).

Experiments that are conducted in a virtual and computational environment often do not allow access to the object of inquiry. The question, whether evidence from *virtual* experiments qualifies as *empirical*, is still debated. One of the issues within this debate concerns the relationship between behavioral models and simulated or virtual objects on one hand, and real world behavior and the real world objects on the other. Are these virtual entities *representations of* or are they *representative for* the real world object?

Computer Models as *Scientific* Experiments

Models representing theories (*conceptual* models) and models representing real entities (*representational* models) must be accommodated within the framework of scientific practice. The conceptual model is the kind of model that has been associated with the terms *metaphor* and *analogy* by Bailer-Jones (2002). The general claim is that *all* models are metaphors. In this view, models are

> an interpretative description of a phenomenon that facilitates access to that phenomenon. […] This access can be perceptual as well as intellectual. […] Models can range from being objects, such as a toy aeroplane, to being theoretical, abstract entities, such as the Standard Model of the structure of matter and its particles (Bailer-Jones, 2002, 108).

Some models can be an adequate representation of real entities provided that there is sufficient accuracy with which a model represents the real world. *Sufficient accuracy* is not a clearly definable term. What is 'sufficient' is essentially a matter of one's subjective stance toward the question what *science* is and how it operates. Psillos, defending the position of scientific realism, says that

> taking a realist attitude toward a particular model is a matter of having evidence warranting the belief that this model gives an accurate representation of an otherwise unknown physical system in all, or in most, causally relevant respects (Psillos, 1999, 144).

We can consider and may even be able to defend the view that models in science go beyond being 'interpretative descriptions' and that they are scientific truths instead. Psillos hints that the adequacy of a model as a representation can only be determined on a case by case analysis, when he refers to the realist attitude toward a *particular* model. We will have to accept that the judgment whether a model or a simulation, or any experiment with such a model, is grounded in some *scientific* method, will also have to be made on a case by case basis. I have already shown that there are no rules for building models, and that the process of building models is largely based on assumptions about what the relevant factors are, how things can be simplified, how we write a program, and so on. The question of whether *virtual* simulations and *virtual* models are valid tools for a *scientific* endeavor is even more problematic. Ziman, who argues for a normative view of science, comments that

> [m]ost people who have thought about this all are aware that the notion of an all-conquering intellectual 'method' *is* just a legend. This legend has been shot full of holes, but they do not know how it can be repaired or replaced. They are full of doubt about past certainties, but full of uncertainty about what they ought now to believe (Ziman, 2000, 2).

I believe that thoughts by Popper (1959) on how science should operate are still normatively useful. Theories should be formulated such that they are testable, and neither magical ingredients nor magical methods should be allowed as part of the supporting evidence. This, of course, must also apply to any counter examples and counter arguments. The application of models is a part of the empirical process. Helping to flesh out the details of some theory or to formulate a new hypothesis using models and simulations is also part of a scientific framework (Popper, 1959,

106). The epistemological role of simulations and models in science in terms of the development of theories is closely linked to questions about scientific theories in general (Peschl and Scheutz, 2001). Nevertheless, I believe that models and simulations are scientific tools, provided 'good scientific practice', whatever that may entail, is applied. Claims for a particular model or simulation should be judged as an adequate representation, and as an adequate, i.e. suitable, model, need to be examined in each case. We need to check that each part and process of a model can be mapped onto the corresponding part of the real world object or process that is modelled. In the case of a computer model, the elements and links in the data structures links and their relations to the object that is modelled need also to be explained. A computer model has to be *testable* in two ways. Firstly, we can test that the model is adequate in terms of what it models, and secondly test how the model is implemented and whether the implementation itself is adequate.

Churchland and Sejnowski (1992) note that real worldliness has two principal aspects, namely (1) that the world is more complex, so that scaling up models does not always succeed, and (2) that real world events do not occur in isolation. Consequently, virtual models and virtual experiments lack realism in several ways. Like many other more 'conventional' models, they do not scale well, both structurally and functionally, and the virtual implementation by means of computation, reduces the number of similarities to real world objects even further. In a way, computer simulations introduce a second layer of abstraction. The first layer is the abstraction or conceptualizing of real world phenomena into a model. The second is the simulation of the model and its dynamics into the realm of the virtual.

Levels of Explanation

Models and simulations are targeted at different levels of explanation. A model can be used to explain certain aspects of a neuron, a particular phenomenon within a neuron, or the behavior of a collection of neurons. Another way to specify levels of explanation of models concerns the model itself. Models and simulations have a high level task to explain something. This level is likely to be the most abstract, and much of the model's implementation and internal workings may be of little interest. If, for example, we are presented with a simulation of the behavior of a few neurons on a computer screen, the actual implementation is of no concern to the observer or experimenter. The neuron simulation works (hopefully) as it should - it should work according to a set of specifications, which the experimenter is aware of. However, there are many layers of programs, library functions, operating system, device drivers, integrated circuits, gates, resistors and wires. The laptop computer, which I am using now, has several quite different programs for neural simulations stored on it. Most of these models are trying to explain the same thing at the highest or abstract level. They are all about relatively simple artificial neural nets, Hebbian learning, learning algorithms e.g. back propagation, and so on. The fact that the 'neurons' in these programs are mathematical structures involving mostly linear algebra is not essential to know or understand in detail for many users of the computer programs. The implementation of the mathematical engine, the subsystem that evaluates and transforms the matrices, is accessible only to the mathematically oriented computer programmer. Then, of course, there are all the components and systems that are part of the implementation on an actual machine. Very few of us have a deep understanding of the technical details of these systems and components.

Conclusion

Models and simulations are part of what is considered *scientific method* in the empirical sciences, although it is not clear what the term *scientific method* actually denotes. In some scientific disciplines, like in the field of Cognitive Science, there are many phenomena which do not belong to the *observable* world; but as Peschl and Scheutz point out that "[i]t is exactly this 'hidden character' of many cognitive processes which makes this domain so interesting as an object of scientific research" (Peschl and Scheutz, 2001).

This holds true for other disciplines. The fact that many processes are not accessible for direct or indirect observation is also interesting in terms of what can be modeled and simulated. It is not so much the mode of experimentation. Whether real world objects or 'virtual objects' are the targets of the experiment does not seem to be that much a point of controversy. It is, I suspect, the human contribution during analysis and interpretation that makes the experiment and the results appear to be 'reasonable' in terms of their value as scientific evidence. We should not forget that with the ever increasing complexity of computer hardware and the operating system software, it is impossible for most application programmers to understand much of these system 'operations' in any detail. Some of the users of software that offers a friendly interface for experimentation with artificial neural nets, for example, may not understand how the neural nets work on a theoretical level, or how they are implemented mathematically or as programs. However this is a point of concern, in the same way it *should be* a concern when using any other kind of technical equipment in scientific experimentation. The challenge remains for the provision of suitable explanations of how the apparatus (computer) works, and more importantly how the model or simulation that is implemented on the computer relates to the real world. The explanations will need to be different, due to the inherent difficulties in demonstrating causal chains in a virtual world.

References

Araya, A. A. 2003. "The hidden side of visualization", *Technè,* 7(2):27–93.

Bailer-Jones, D. M. 2002. "Models, Metaphors and Analogies", in: P. Machamer and M. Silberstein, eds., *The Blackwell Guide to Philosophy of Science,* Malden: Blackwell, 108–127.

Churchland, P. S. and Sejnowski, T. J. 1992. *The Computational Brain*, Cambridge: MIT Press.

Dretske, F. 1988. "Representational Systems", in: T.O'Connor and D. Robb, eds., *Philosophy of Mind: Contemporary Readings*, New York: Routledge, 304–331.

Gooding, D. 2003. "Varying the Cognitive Span: Experimentation, Visualization, and Computation", in: H. Radder, ed., *The Philosophy of Scientific Experimentation*, Pittsburgh: UoPP, 255–283.

Green, C. D. 2001. "Scientific models, connectionist networks, and cognitive science", *Theory & Psychology*, 11(1):97–117.

Hacking, I. 1983. *Representing and Intervening*, Cambridge: Cambridge UP.

Harré, R. 1970. *The Principles of Scientific Thinking*, Chicago: UoCP.

Harré, R. 2003. "The Materiality of Instruments in a Metaphysics for Experiments", in: H. Radder, ed., *The Philosophy of Scientific Experimentation*, Pittsburgh: UoPP, 19–59.

Jordanova, L. 2004. "Material Models as Visual Culture," in: S. De Chadarevian and N. Hopwood, eds., *Models: The Third Dimension Of Science*, Stanford: Stanford UP, 443–451.

Jorion, P. 1999. "What do mathematicians teach us about the world? An anthropological perspective", *Dialectical Anthropology*, 24(1):45–98 (reprint 1–25).

Keller, E. F. 2003. "Models, Simulation, and 'Computer Experiments'", in: H. Radder, ed., *The Philosophy of Scientific Experimentation*, Pittsburgh: UoPP, 198–215.

Krebs, P. R. 2005. "Models of cognition: Neural possibility does not indicate neural plausibility", in: B. Bara et al., eds., *Proceedings of the 27th annual meeting of the Cognitive Science Society*, Mahwah: Lawrence Erlbaum Associates, 1184–1189.

Krebs, P. R. 2007. "Smoke Without Fire: What do virtual experiments in cognitive science really tell us?", in: N. Srinivasan et al., eds., *Advances in Cognitive Science*, Delhi: Sage. (in press)

Lenat, D. B. and Guha, R. V. 1990. *Building Large Knowledge-based Systems*, Reading: Addison-Wesley.

McClelland, J. and Rumelhart, D. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Volume 2: *Psychological and Biological Models*), Cambridge: MIT Press.

Morgan, M. S. 2003. "Experiments without Material Intervention", in: H. Radder, ed., *The Philosophy of Scientific Experimentation*, Pittsburgh: UoPP, 216–235.

Peschl, M. E. and Scheutz, M. 2001. "Explicating the epistemological role of simulation in the development of theories of cognition", *Proceedings of the seventh colloquium on Cognitive Science* ICCS-01, 274–280.

Popper, K. R. 1959. *The Logic of Scientific Discovery*, London: Routledge.

Psillos, S. 1999. *Scientific Realism: How science tracks truth*, London: Routledge.

Rumelhart, D. and McClelland, J. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Volume 1: *Foundations*), Cambridge: MIT Press.

Russell, B. 1948. *Human Knowledge: Its Scope and Limits*, London: Allen & Unwin.

Stufflebeam, R. S. 1998. "Representation and Computation", in: W. Bechtel, W. and G. Graham, eds., *A Companion to Cognitive Science*, Malden: Blackwell, 636–648.

Ziman, J. 2000. *Real Science: What it is, and what it means*, Cambridge: Cambridge UP.

Reasoning with Safety Factor Rules

Jonas Clausen and John Cantwell
Division of Philosophy
Royal Institute of Technology
Stockholm,Sweden

Abstract
Safety factor rules are used for drawing putatively reasonable conclusions from incomplete datasets. The paper attempts to provide answers to four questions: "How are safety factors used?", "When are safety factors used?", "Why are safety used?" and "How do safety factor rules relate to decision theory?". The authors conclude that safety factor rules should be regarded as decision methods rather than as criteria of rightness and that they can be used in both practical and theoretical reasoning. Simplicity of application and inability or unwillingness to defer judgment appear to be important factors in explaining why the rules are used.

Keywords: Safety factor, uncertainty factor, decision theory, reasoning, heuristics

Introduction

I'm driving along in my car, and it's a beautiful day. In front of me on the highway is another car, and as I'm driving rather fast I'm closing quickly. Then I remember the "Three Second Rule". The rule says that, when driving on the highway, I ought to stay at least 3 seconds behind the car in front for reasons of safety. I slow down and start counting the time between us, and after a few moments I have adjusted the speed and distance so that I'm just over 3 seconds behind the car in front. I relax, as I am now confident that I am driving in a sensible and reasonably safe manner in line with good driving practice.

The "Three Second Rule" is an example of a decision rule, or a decision *heuristic*, that contains a *safety factor*, in this case: three seconds. The use of safety factors is widespread. In civil engineering time-tested multipliers from certain key load values (*e.g.* estimates of average or maximal load) to reasonable design strengths serve as heuristics for safe construction. In toxicology there are simple heuristics for drawing reasonable conclusions from incomplete datasets regarding chemical effects on humans. These rules make use of *uncertainty factors* or safety factors as divisors from, say, results obtained in mice to putatively reasonable estimates in humans.

In decision theory rules of thumb have traditionally been regarded with mild suspicion: they have been treated as objects not quite worthy of serious theorizing. Good decision-making should be based on numerical probabilities and utilities, or at least be reducible to these concepts. In the last decade or so, with the advent of computer assisted, and automatic, decision making, this picture has changed. Resource bounded rationality has become a field of its own, and rules of thumb form an important sub-class of decision methods.

In this context decision making with safety factors is a curious hybrid. On the one hand we have the "Three Second Rule" in traffic, an unsophisticated but helpful guide to action. On the other hand we have the systematic application of safety factors in various fields of engineering and in

fields like toxicology. In these areas the safety factors involved have been put under close scrutiny, both as regards the proper calibration of the numbers employed in the safety factor and as regards their theoretical status.

In this paper the status of *safety factor rules* as a decision making tool will be scrutinized. How are they used? Why? When? And what is their place in decision theory? The structure of the paper is as follows. In Section 2 two uses of safety factor rules are presented in more detail (still, rather schematically). One example is taken from toxicology, the other from the design of structural components. In Section 3 we try to place safety factor rules within a larger context of decision theory, somewhat hesitantly identifying safety factor reasoning with a form of 'satisficing'. Section 4 addresses the question whether safety factor rules should be seen a part of the process that risk researchers call 'risk assessment' or if it should be seen as part of the process of 'risk management'. It turns out that the answer varies and that the use of safety factor rules often makes the distinction difficult to uphold. In Section 5 we discuss the question of safety factor rule justification and the trade-off made in science policy decisions. Conclusions are then presented in Section 6.

Examples of safety factor rules

Partial safety factor method for structural design

A frequently used method for designing structures is the *partial coefficient method* or *partial safety factor method*. This can be formulated in several different ways of which the following is one. Assume that the failure propensity of the structure is governed by load type variables S_i and resistance type variables R_j. The safe set of the structure is then assumed given by

$$g(S_i, R_j) \geq 0$$

Partial safety factors γ_{Si} and γ_{Rj} are numbers equal or larger than unity. The safety margins are considered adequate provided that

$$g(S_i\gamma_{Si}, R_j/\gamma_{Rj}) \geq 0$$

The arguments of the *g*-function are termed design variables. The variables S_i and R_j are most often chosen as characteristic values, S_{ik} and R_{jk} A characteristic values is normally a quantile, such as 0.05 for resistance type variables or 0.95 for load type variables, of the stochastic distribution connected to the variable in question.. The use of characteristic values is recognized as being a vast simplification (Ditlevsen and Madsen, 2004, p. 22), since it amounts to representing a stochastic variable by one or a few values.

Although actual regulations are in general more complicated it will suffice for our purposes to look at a one-dimensional variant of the safety factor rules used.

Simplified Partial Safety Factor Rule: $S_i\gamma_{Si} \leq R_j/\gamma_{Rj}$

Subfactors that can enter into γ_S and γ_R are factors representing measurement or model uncertainty and so-called *safety classes*, meaning to what extent humans will be in or around the structure. (cf. Boverket, 2003)

An uncertainty (safety) factor rule for human health risk assessment

In toxicology, uncertainty factors are used when making inferences from animal data about dose/response to a reference dose (RfD)[1]. An RfD is commonly understood as "intended to identify a dose or exposure unlikely to put humans at appreciable risk" (Brand *et* al., 1999, p. 295). Starting off with a key dose value such as an animal bioassay NOAEL (no observed adverse effect level) or BMD (benchmark dose) for a certain effect (often called *endpoint*), one then divides it by the safety factor U and the result is the RfD. This is a common rendering of an uncertainty factor *definition*:

$$RfD = \frac{NOAEL}{U_A \times U_H \times U_S \times U_D \times M}$$ (Gaylor and Kodell, 2000)

The different divisors are explained as follows:

- U_A is the *interspecies factor* for using animal data for human response, say from studies on mice. A common value is 10 (Dourson *et* al., 1996).
- U_H is the *intraspecies factor* for considering sensitive subgroups in the general human population, such as pregnant women or people with inherited susceptibility to certain substances. Again, a common value is 10, though the factor is at times as low as 1 (Dourson *et* al., 1996).
- U_S is the *chronicity factor* for using subacute (very short-time) or subchronic (short-time) data for chronic (long-time) effects. Subacute studies are normally conducted over 14 days, subchronic studies over 90 days and chronic studies over approximately 2 years (Kalberlah *et* al., 2002). Values less than 10 are normally used (Dourson *et* al., 1996)
- U_D is the *database factor* for using incomplete data sets, such as studies that do not cover enough of the possible adverse effects. Values for U_D vary from 1 to 100 (Dourson *et* al., 1996).
- M is the *modifying factor* to be used for further considerations according to expert judgment and is normally less than 10 (Dourson *et* al., 1996).

Also, Burin and Saunders (1999) note the following:

> The uncertainty factors usually range from 1 to 10 depending on the extent of the uncertainty. As uncertainty is reduced, a smaller factor may be used. (p. 210)

Although the former certainly seems true, the latter is not always the case. Even if a factor of 10 is often the default choice when uncertainty is very large, it is a clear possibility that less uncertainty regarding, say, the relation between sensitivity of rats and humans could warrant a larger interspecies factor than 10 if the information obtained indicated that the substance had effects to which humans were much more sensitive than rats.

An interesting thing about this uncertainty factor definition is that the right side of the equation is available to a risk assessor through a fairly well defined procedure. The NOAEL can be obtained through routine testing and the division of that result by the factors is a simple mathematical procedure. One might then say that the relationship *operationalizes* the RfD.

The safety factor rule based on the above definition and the common understanding of the RfD is, we would argue, something along the following lines:

> **NOAEL Rule:** *A dose less than NOAEL/U is unlikely to put humans at appreciable risk*

The NOAEL could of course be exchanged for a BMD for an analogous BMD Rule.

Although something like this can be hard to find stated explicitly, it is hard not to interpret the terms in this way. In fact, the NOAEL Rule we have suggested is an implication of the safety factor definition (NOAEL divided by U gives RfD), the definition of the RfD (being a dose that is relatively "safe") and a *monotonicity assumption*;[2] that a smaller dose will always mean less or equal probability of a certain response than a larger dose, and will thus be safer.

Decision theory and safety factor rules

When driving we want to avoid accidents, when building we want to avoid that the structures collapse, the overriding goal of toxicology is to establish at what dosage a substances poses a health-threat to humans. We use safety factors to be on the 'safe' side. A number of questions arise.

Why take a perhaps costly measure to be on the safe side? Why not simply follow the course of action that strikes an optimal balance between the values involved (travel time, building cost, risks to human health, etc.)? Classical decision theory tells us to do just this. It claims that an action is rational if it has the highest *expectation value* of all alternative courses of action, where the expectation value can be expressed by (the o_i :s are the possible outcomes of A, $Pr_A(o_i)$ is the probability that A will have the outcome o_i, and $V(o_i)$ is the value of the outcome) :

$$EV(A) = \sum Pr_A(o_i) \cdot V(o_i)$$

A major problem is that typically we have only a vague appreciation of the probabilities involved in a decision problem, and that a good, non-arbitrary, numerical estimate of the values involved is hard to come by. A second standard criticism is that as a psychological fact we seldom if ever compute probabilities and values in the way prescribed by the expectation value approach, and the very act of computing them would, in some situations, be harmful as it would distract our attention from the situation at hand.

So, finding an optimal balance requires that the different values involved are fully comparable, that the probabilities of adverse outcomes are known (even though they be costly or unethical to acquire), and that we have the time, attention, and money to engage in the activity of optimizing. The three second rule is easy to apply and lets me keep my attention on my driving. We just cannot establish the dose-response curve for a chemical by testing it on humans because of ethical constraints on research. Built structures have so many interrelated components and can be subjected to so many different and varied kinds of external forces that only highly sophisticated computer models can even begin to take in the complexities. These are reasons why the principle of maximizing expected utility is of limited practical use in many areas, but a nagging question remains: are safety factors a good replacement for optimizing?

We must keep in mind what we mean by a 'replacement'. The principle of maximizing expected value (MEV) or maximizing expected utility (MEU) can be viewed in two ways. On the one hand it can be seen as giving a *decision method*, an algorithm by which one deduces which action to perform. However, many decision theorists and philosophers that endorse the principle, view it not primarily as a decision method, but as giving a *criterion of rightness*. A rational person should act so as to maximize expected utility. This is not the same as saying that a rational person should *calculate* the expected utility before acting, indeed in many cases sitting down to perform a number of calculations would be the *wrong* thing to do. Rather you should act *as if* you had done the necessary calculations.

Isaac Levi (1981) has developed a decision theory that, even as a criterion of rightness, relaxes the constraint imposed by classical decision theory. Instead of basing a decision on a *single* probability function and a *single* utility function, Levi's theory allows the rational agent to have sets of probability functions and sets of utility functions, and rational decisions are characterized in terms of these sets.

Levi's decision criteria are *lexicographic*. First select those actions that maximize expected utility according to *some* combination of probability function (taken from the set of probability functions) and utility function (taken from the set of utility functions). If several actions satisfy this constraint, select that action that maximizes the minimal expected utility (the minimal value we get from some probability function and some utility function).

Satisficing

The idea of *satisficing* was first introduced by Simon as an alternative to classical decision theory. It can be interpreted both as *criterion of rightness* and *decision method*, and it can be applied in two different decision phases: choice and pre-choice deliberation.[3] For the choice phase the idea of satisficing can be formulated:

> **Alternative satisficing** (Decision rule interpretation): An alternative is rational iff it has (expected) value that equals or exceeds the aspiration level **α**.

This is one interpretation of the discussion of *procedural rationality* in Simon (1976). The aspiration level **α** here tells us when an (expected) outcome is "good enough" or "satisfactory".

The idea of alternative satisficing as a criterion of rightness has been severely criticized. How can it be rational to perform an action that is "good enough" if one knows that there is an alternative that has a better outcome? It has been convincingly argued by Richardson (2004) that this idea is incoherent. In brief, the argument goes that either the concept of value needed for alternative satisficing to work cannot be made sense of or satisficing is uninteresting as a concept. One alternative is that value is of the "all things considered" kind, and then doing something that one recognizes as worse "all things considered" than some other available option, something allowed by alternative satisficing, is simply not intelligible as rational behavior. If value is not of this kind, then "satisficing will merge indiscriminately with the simple and banal idea of tradeoffs." (ibid., p. 108).

Alternative satisficing as a decision method can also be criticized on the same grounds as MEU; that it is, in certain cases, equally impossible for a non-ideal agent to find a satisficing alternative

as it is to find an alternative that maximizes expected utility (given extraneous utilities). For example, the agent must in the worst case (only one satisficing option and it is found last, if at all) examine all possible options and compare with the aspiration level, and this task might indeed be intractable.

Taken together, these lines of criticism present a serious challenge to alternative satisficing both as decision method and criterion of rightness.

Deliberation satisficing tells us when to stop our pre-choice search for alternatives and proceed to actually choosing an alternative. This is the "stop rule" or search rule interpretation and can be stated:

> **Deliberation satisficing** (stop rule interpretation): The search for further alternatives can stop iff (at least) one alternative with (expected) value at or above α has been found.

This is an interpretation of the discussion of stop rules in Simon (1972). Deliberation satisficing understood as a decision method for the "pre-choice choice" or meta choice to search or not, allows it to be made without evaluating search branches. Only currently available alternatives need to be evaluated when deciding whether to look for more. This is of course a potentially huge computational saving, but how well it works depends on how close the aspiration level is to the actual optimum (if there is indeed a well-defined optimum), and it will have the same worst-case characteristics as alternative satisficing. It should be noted that deliberation satisficing says nothing whatsoever about the search process itself, only about when it should begin and end.

Deliberation satisficing can also be seen as a criterion of rightness, and tells us when it is rational to keep on or cease searching, and this is a question that is answered with reference to the values of available alternatives. Again, the question arises of why we should stop the search at some suboptimal point if we know there are better ones (in the sense of all-inclusive value), and the same criticism that was voiced against alternative satisficing as a criterion of rightness can be directed against deliberation satisficing.

An important variation of the stop rule is to relativize it to a particular parameter. For instance, once we find that a particular drug is "safe enough" we can stop looking for safer alternatives, and instead direct our attention to making the production of the drug cheaper. On this interpretation the aspiration level is set not on the combined result (the amalgamation of the different parameters), but on different parameters. This 'parameterized' stop rule is of particular interest in settings where diverse values that are difficult to compare are at stake (such as health vs. cost), or where we have good reason to believe that we know some upper limit or approximate optimum in some dimensions but not in others.

Safety factor rules, maximization and satisficing

Consider again the three second rule. It is based on a single, easily obtainable parameter: how many seconds ahead the next car lies. It encapsulates two opposing values: the value of getting to your destination quickly and the negative value of running into the car ahead. It also encapsulates a certain amount of probabilistic information: with a three second safety margin, chances are that if the car ahead slows down quickly, or stops, you will be able to stop without running into it. Part of this probabilistic information is based on knowledge of reaction times, and of the functionality of typical brakes. Thus, for all its simplicity the three second rule encapsulates both

the values we ascribe to quick and safe transportation, and considerable knowledge about the behavior of humans and cars.

How have all these different features been combined so as to result in the three second rule, rather than in the four second rule, the two second rule or the 2.9999 second rule? Obviously, the 2.9999 second rule would be dismissed on the basis of being difficult to use. What about the two second rule? Here one can probably argue, and show, that it gives too little margin for people's widely varying reaction times. The four second rule, on the other hand, could be rejected on the basis that the three second rule provides enough safety anyhow: it is satisficing with respect to safety.

So the three second rule is not taken out of a hat. It can (possibly) be reconstructed as being based on a reasoned weighing of values against probabilities. But, of course, it is still far from being derived using the principle of maximizing expected utility, either in its classical form or in the relaxed version given by Levi. If anything, the three second rule has been derived from practice. One could hope, perhaps in this case even suspect, that with careful numerical estimates of the values and probabilities involved, we could derive the three second rule, but, in this case at least, such an analysis seems pointless: the rule works well enough.

In much of this the three second rule has features similar to those of safety factor rules used in engineering and toxicology. In these areas too, the safety factors encapsulate both values at stake and specific knowledge about the processes involved. In these areas too, the safety factor chosen is taken to give a 'big enough' safety margin (satisficing with respect to safety) and endeavor to smooth out individual differences in specific materials and humans. One would suspect, however, that in these areas we would not accept the cavalier attitude that the safety factors are not in need of further analysis on the grounds that they 'work', for our impression that they work can be based on scarce information. And indeed it is part of the praxis of these disciplines to refine the grounds from which safety factors are derived. But lack of information will always be a problem and to some extent the safety factors have been chosen because they 'work'.

To conclude this section, superficially, safety factor rules appear quite far away from the paradigm of using MEU (or Levi's variation) as a decision method. However a closer analysis shows that they encapsulate both probabilistic and value-based information, but encode a satisficing element with regards to safety.

Practical and theoretical reasoning in risk assessment and risk management

The standard model of the relation between risk assessment and risk management is sometimes simply called the *risk assessment/management paradigm*. The model is temporally ordered in the sense that research must be concluded (insofar as research can be concluded) before assessment can conclude, and assessment concluded before management. However, initiation of e.g. the management subprocess will at times be first in the chain of events, so the order of initiation is not as clear. Questions for which there are no readily available answers are passed to the left in the model and answers returned to the right (see Fig I).

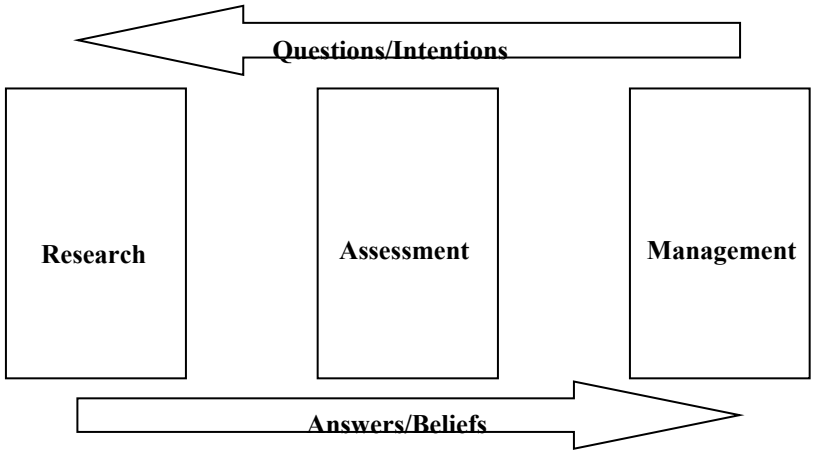

Figure Flow of intentions and beliefs in the risk assessment/management paradigm.

Although research, assessment and management each can happen more or less independently, the leftmost subprocesses can be "nested" in a process to the right – research can be nested in the assessment process and assessment nested in the management process. The nesting is, to put it simply, the result of questions flowing left in the model and answers flowing right.

"Science and Judgment in Risk Assessment" (National Research Council, 1996) describes *risk assessment* of chemicals as follows:

> Human-health risk assessment entails the evaluation of information on the hazardous properties of environmental agents and on the extent of human exposure to those agents. The product of the evaluation is a statement regarding the probability that populations so exposed will be harmed, and to what degree. The probability may be expressed quantitatively or in relatively qualitative ways. (pp. 25-26)

While risk assessment is often thought of as science-based, risk management involves further considerations. Just as with risk assessment, "Science and Judgment in Risk Assessment" contains a description of *risk management* of chemicals:

> Risk management is the term used to describe the process by which risk assessment results are integrated with other information to make decisions about the need for, method of, and extent of risk reduction. Policy considerations derived largely from statutory requirements dictate the extent to which other factors – such as technical feasibility, cost and offsetting benefits – play a role. (p.28)

Gaylor and Kodell (2002) distinguish between safety factors that are *risk reduction factors* and those that are just estimations of quotas between the dose-response curves for different populations (animals/humans or human population in general/sensitive subpopulations). This can

be interpreted as a distinction between risk management factors and risk assessment factors. The nature of this distinction will be the topic of part 4.3.

The distinction between risk management and risk assessment superficially parallels the distinction between *practical reasoning* and *theoretical reasoning*. Both practical and theoretical reasoning are distinguished by their end results. A piece of practical reasoning is a reasoning process that ends in action or, more plausibly, intention.[4] Theoretical reasoning on the other hand ends in belief. There is also what John Broome (2002) calls *normative reasoning*, which amounts to theoretical reasoning about normative propositions.

In spite of the similarities in end results, with both risk assessment and theoretical reasoning leading to beliefs and practical reasoning and risk management leading to intentions to act, it is not the case that risk assessment *is* theoretical reasoning, nor that risk management *is* practical reasoning. The reason for this is that neither risk assessment nor risk management consists only of reasoning. Further, there are practical and theoretical reasoning processes involved in both risk assessment and risk management.

The normativity of safety factor rules

The calculation of an RfD, as in 2.2, is normally regarded as risk assessment. This means that, according to the received view, it is supposed to be a scientific or at least science-based, and as such non-normative. One understanding of the safety factor rule is empirical. The NOAEL or BMD values are results from rather straightforward experimental procedures. If we consider the RfD to be a non-normative concept, then the uncertainty factor rule is relatively innocuous, as it is not to be understood as action guiding. It merely describes a manner of using words. However, it certainly seems as if an element of normativity has snuck into the idea of an RfD, as can be seen in the quote given earlier: "…*unlikely* [our emphasis] to put humans at *appreciable* [our emphasis] risk.". It is arguably a normative issue what we consider to be unlikely,[5] not to mention when a risk is appreciable, since it appeals to the intuition that we need not care about unlikely or unappreciable risks. So, if the RfD is interpreted normatively, we have something that isn't quite as innocuous, namely the claim that finding a certain experimental value and dividing it by suitable factors presents us with a dose that is at least *prima facie* nothing to worry about.

The safety factor rule in 2.2 is more openly normative since it speaks of design values which according to Ditlevsen and Madsen (2004) should be interpreted in such a way that a structure is "just sufficiently safe" if it is constructed using design values (Ditlevsen and Madsen, 2004, p.22). To build a structure with parameters implying safety beyond that provided by building with design values is, according to such a view, going over and above what can reasonably be required. It is *supererogatory* if you will.

The normativity of safety factor rules makes them controversial, but it is also the very thing that makes them useful, not only in practical reasoning during the risk management and assessment processes, but also in normative reasoning about the results of risk management and assessment.

Practical and theoretical reasoning with safety factor rules

The following is a "just so" account of the role played by safety factor rules in toxicological risk assessment and structural engineering. With "just so" we mean to say that the account should not necessarily be taken as an empirical conjecture. It is more of a demonstration of possibility,

showing how safety factor rules and definitions *can* be used in inferences. This is sufficient for answering the question of whether we can make sense of the distinction between assessment safety factor rules and management safety factor rules.

Reasoning with the toxicological safety (uncertainty) factor rule

To recap, the safety factor rule mentioned in 2.2 was the following:

> **NOAEL Rule:** *A dose less than NOAEL/U is unlikely to put humans at appreciable risk*

In toxicological risk assessment, the "just so" story starts out with an intention to find an RfD, or a dose unlikely to put humans at appreciable risk. The NOAEL Rule tells us that a sufficient means to finding such a dose is to find a NOAEL and divide by a suitable U. This corresponds to a "leftward" motion in the research-assessment-management model, from a question belonging to risk assessment to a question for research – to find a NOAEL.

However, the motion for which the rule can be used is also a "rightward" one. When an answer has been provided by research, such as the specific value of a NOAEL, we can use the NOAEL Rule to infer an RfD, by dividing the NOAEL by U.

The first reasoning that makes use of the rule is a piece of practical reasoning from an intention to find out something necessary for risk assessment to an intention to do certain research. The second piece of reasoning is theoretical and takes us from a research answer to a risk assessment answer. Since both these pieces of reasoning can be nested within a risk management process it could be argued that in such a nested case, any safety factor used is possibly done so, in a sense, in an encapsulating risk management process.

Reasoning with the engineering safety factor rule

As above, we will recap the safety factor rule mentioned earlier:

> **Simplified Partial Safety Factor Rule:** $S_i\gamma_{Si} \leq R_j/\gamma_{Rj}$

While risk assessment in toxicology is about finding "safe" doses, risk assessment in structural engineering can be seen as finding "safe" designs or evaluating designs with respect to safety. Just as the NOAEL Rule in 4.2.1, the Simplified Partial Safety Factor Rule, with given safety factors, tells us that if we want to find a sufficiently safe design we need to find characteristic values (reasoning "leftwards" from intentions for risk assessment to intentions for research). And, as above, the other direction of reasoning is possible when we are faced with a structure with certain characteristic values for materials given from research. We can then infer whether the structure is safe or unsafe by calculating the "implied" safety factor and compare it to our code.

Comments

The safety factor rules can arguably be used in both assessment and management because of the nesting of management, assessment and research processes, as well as the dual "directions" of reasoning made possible by the rules. Thus, a distinction between management and assessment safety factor rules and definitions does not lie in when they *can* be used. Might it lie instead in when they *should* be used? Again, nesting and dual use present problems, for say e.g. that we

create a compound factor, multiplying all the needed factors – be they considered assessment factors, management factors or other – that will takes us from a research result to a risk management decision. Is the calculation with such a compound factor to be seen as management, assessment or what? If the subfactors are justified for use separately, surely it will be justified to use a compound factor that is not easily identified as either an assessment or management factor. A remaining possibility is that assessment safety factor rules *are* used only during the risk assessment phase, and that management safety factor rules are used only during the risk management phase, but the plausibility of the dual directions of reasoning seem to speak against this. Further, nesting again presents a problem of placing a certain event squarely in only one of the research, assessment and management categories.

One possible reaction to the accounts of 4.2.1 and 4.2.2 is that they are simple, and perhaps too simple. This is, we would argue, precisely the point. Our conjecture is that the use of safety factor rules owes much to the simplicity of the reasoning involved. What can be added at this point is that a simple rule with simple reasoning can fail to do what it is supposed to do, and that a far more complex rule might do the job better,[6] if the job is understood as enabling accurate conclusions. Banal as it may seem, safety factor rules are often a trade-off between simplicity of reasoning and accuracy of conclusions, with the prime difficulty being that we cannot normally say *how exactly* the trade-off looks.

Discussion

Valid inferences and correct results

As mentioned in 4.2, safety factor rules play the role of "bridge hypotheses" and are the answer to science policy issues (RIAP, 1994). They enable agents who believe in them to make valid inferences about important issues, where valid is to be understood as logically valid. Logically valid inferences, however, do of course not guarantee that the conclusions derived are correct. Take the racist inference rule of "If someone has a different skin color than you do, that person is unintelligent". Conclusions derived about the intelligence of others with the help of this rule may be logically valid, although they will often be inaccurate.

Do safety factor rules go too far?

Commonly used safety factor rules are generally not thought of as necessary for safety, but rather as sufficient since they are often thought of as conservative or cautious. This suggests that if we knew more we could act with lower regard for the safety factor rule, and still be safe. One criticism that can be voiced against the use of safety factor rules relates to this, and it is that they enable unwarranted conclusions and might thus not, in fact, be sufficient for safety. In a choice between using a safety factor rule and statistical methods, one can ask what conclusions will be enabled by each approach. Let us assume that we are doing measurements on rats examining the prevalence of blindness resulting from exposure to some substance S. The results from the study indicate that the NOAEL is 4 mg/kg bw and that the highest dose not provoking blindness at 0.95 confidence level is the range 2-7.3 mg/kg bw. A further study on the general chemical sensitivity of rats as compared to humans gives, say, that humans are 0.22 - 13 times more sensitive per kg bw at 0.95 confidence. This gives us a range from 0.148 to 33.2 mg/kg bw for the highest dose not provoking blindness at confidence ≥ 0.9.[7] The result from the default uncertainty factor rule is that a dose under 0.4 mg /kg bw is safe using an interspecies factor of 10. Now, although this example is entirely fictional and many details have been omitted, we would argue that this is how

different the conclusions from the statistical approach and the safety factor approach can be, even given the same information. In fact, the less statistical information we have, the more divergent results will be since intervals with a fixed confidence level will become larger.

If the conclusions made possible by safety factor rules outstrip those made possible by statistical methods, those "extra" conclusions (such as altering the "safe" interval) will be unwarranted according to the confidence level chosen for the statistical analysis. One could, for example, experiment with confidence levels to see at which point the conclusions warranted by a safety factor rule become warranted by the statistical analysis. In this regard it cannot be the case that more "allowing" safety factor rules can replace statistical analysis without a loss in epistemic reliability, that is, without implying larger epistemological risks. Obviously, if we chose to have a very low confidence level in the statistical analysis virtually any conclusion can be "supported" by the analysis.

Evolution of safety factors

There is an interesting difference between safety factor rules that have a long history and statistical analyses based on more recent studies or compilations of data, which might affect how safety rules unsupported of statistics are viewed. It is a difference similar to that between *batch* and *online learning* in Computer Science. Online learning is myopic in that it gives incremental output to incremental input, while batch learning takes into account all the available input.[8] To see how they are different, imagine that you have one hour to find the highest point you can in a hilly landscape. Before beginning the task you are blindfolded, so the only way of finding your way is by moving around the landscape sensing the incline. If you wanted to solve the problem in an "online" way you would at each point in your "optimizing walk" decide where to go next and hope that that next step would take you to the highest point, and after each step forget all about where you had been previously. After one hour you simply stop. Solving it in a more "batch"-like way would be to first walk around a while, collecting data by memorizing the entire walk, and then try to infer where the highest point lies. Batch processing requires more memory, namely enough for the entire sequence of input, while online processing has the downside of not being repeatable or open to scrutiny unless the same sequence of input is presented again.[9]

In a similar way, we can see that safety factor rules in some areas have been around for a long time, and some of these factors have been incrementally changed over time, presumably as reactions to events related to their use or new research results. The values they do have may be supported by good reason, although the details of those reasons are sometimes lost. Thinking along these lines relates to Ditlevsen's (1997) discussion about a "superior authority" within a country or union of countries that, even in cases where codes have not been calibrated using modern statistical methods, decides what designs or codes are to be considered optimal. An interesting question then becomes how good statistical data we need before deciding to alter an "online" safety factor rule in a batch-like fashion. It is not a question to which we have an answer, but it suffices to acknowledge for the moment that it presents a serious complication for normative evaluation of safety factors that are not supported by readily available statistical data.

Safety factor rules are responses to science policy issues

The terms *science policy issues* and *science policy decisions* can be used to further explain safety factor rules. Here we need to distinguish between provable and unprovable risks. The following quote from Choices in Risk Assessment (1994) gives a characterization:

> Provable risks can be measured or observed directly and include actuarial risks such as those associated with highway or air travel accidents. In contrast, other risks – such as those associated with low-doses of radiation or exposure to chemicals in the environment – are often too small to be measured or observed directly with existing scientific methods and available resources. Additionally, specific health and environmental effects are often difficult to attribute to specific causes because other competing causes cannot be excluded with reasonable certainty. Such risks are unprovable. (p. 241)

The next quote gives the definition of science policy issues and decisions:

> When risk assessment is used to estimate unprovable risks, these gaps and uncertainties [in scientific knowledge, data and method] become science policy issues. Both risk assessors and risk managers make science policy decisions in order to bridge the gaps and uncertainties. Thus, science policy decisions enable the estimation of unprovable risks. (ibid, p. 241)

Even though Choices in Risk Assessment focuses on chemical risks, the idea is quite general. In other words, when we lack solid information, we have to make educated guesses, *given that we must provide an answer*. In the face of uncertainty we have two basic ways to go: defer judgment or guess. Science policy questions are questions of how we ought to guess under the difficult circumstances mentioned, given that deferring judgment is out of the question. Such guesses do not come without a cost (of sorts). Whatever answer we provide will have a less than ideal (or as is normally the case, less than scientific) reliability, and acting upon it means taking what Sahlin and Persson (1994) call an *epistemic risk*. This does not imply that we are taking an *outcome risk* (doing something that has possible unwanted outcomes) of a certain magnitude, but it does imply that we are uncertain about the magnitude of outcome risk we are in fact running. Thus, recognizing that safety factor rules are, in many cases, responses to science policy issues tells us that they are not standards with which we can rest easy.

Conclusions

The questions we set out to answer were "How are safety factor rules used?", "Why are they used?", "When are they used?" and "What is their place in decision theory?".

The answer to the first of these is that safety factor rules are used in at least two forms of reasoning: (i) "leftwards" practical reasoning about sufficient information gathering given needs in risk management and risk assessment towards research and (ii) "rightwards" theoretical reasoning in the direction from research results, through risk assessment results to risk management decisions. That the same rule can be used for both these forms of reasoning was presented as one of two arguments against dividing safety factors into "risk reduction factors" (or management factors) and assessment factors, the other being the possible "nesting" of research, assessment and management processes.

Concerning why safety factors are used, there are several different explanations. One is simply that in certain situations of radical uncertainty we have made a meta decision that we nevertheless must provide answers to certain questions, such as "Is this structure safe?" or "Is this dose safe?", and that reliance on either statistical data or the evolutionary process that produced a certain safety factor rule is strong enough. We have, often implicitly, deemed the epistemic risk inherent in using the safety factor rule acceptable. For other situations, where more precise calculations *can* be made, using safety factors is a simple alternative, a heuristic, and when carefully chosen the safety factor rule can be equivalent, or approximately equivalent, to more complex procedures for a suitably restricted range of cases. In certain cases, when safety factor rules are used unreflectively, one may of course say that they are used because of tradition or simply because regulations force us to, since their use is at times mandatory.

The "When" question has been answered, at least partially, but something can be added. The situations in which the rules are used are situations of varying degrees of uncertainty. Were there no uncertainty, safety factor rules would be superfluous. However, just uncertainty is not enough to motivate their use. The agents using safety factors are generally resource constrained. Safety factor rules allow for resource-bounded decisions to be made systematically, making behavior, at least in principle, open to deliberate revision.

Finally, when it comes to their relation to decision theory, safety factor rules should be seen more as decision methods, tightly connected to highly specific circumstances such as "driving on the highway" or "designing a structural component", than as criteria of rightness. They encode trade-offs between various values at stake and beliefs about the world, but on a superficial level they are satisficing with respect to safety, in the sense that they tell us when something is "sufficiently safe". However, since the precise formulation of a safety factor rule is often a matter of science policy decision-making, this tells us that any such statement of sufficient safety is provisional.

Acknowledgments

This paper is part of a project financed by the Swedish Research Council. The authors would like to thank Fred Nilsson and Sven Ove Hansson for valuable comments on earlier version of this paper.

Notes

1 Several other key dose values are or have been in use. Among these we find *tolerable daily intake* (TDI), *acceptable daily intake* (ADI) and *provisional tolerable weekly intake* (PWTI) (Herrman and Younes, 1999).

2 This assumption is not uncontroversial. There is a discussion in toxicology about the nature of *hormesis*; when a substance gives rise to higher rates of a certain adverse effect at a low dose than it does at a higher dose. See, for example, Calabrese *et* al. (1999).

3 The distinction between decision phases in this way is Simon's (1977).

4 "Intending to act is as close to acting as reasoning alone can get us, so we should take practical reasoning to be reasoning that concludes in an intention." (p. 1, Broome, 2002)

5 Is probability 0.5 unlikely? Is 10^{-3}? Or need we go as far as 10^{-6}? There is serious vagueness here and thus ample room for a broad range of values to affect interpretation.

6 The work of Gigerenzer and Todd suggests that simple rules may in fact do very well under suitable circumstances. See, for example, Gigerenzer and Todd (1999).

7 According to Bonferroni's inequality.

8 For a discussion of batch and online learning, see for example Barkai *et* al. (1995).

9 Many algorithms can be formulated in equivalent variants, either online or batch. This equivalence is in terms of eventual results, not in such things as memory requirements or execution speed.

References

Barkai, N., Seung, H.S. and Sompolinsky, H. 1995, "Local and Global Convergence of On-Line Learning", *Physical Review Letters* 75(7):1415-1418.

Boverket 2003, *Regelsamling för konstruktion – Boverkets konstruktionsregler, BKR byggnadsverkslagen och byggnadsverksförordningen*, Boverket.

Brand, K.P., Rhomberg, L. and Evans, J.S. 1999, "Estimating noncancer uncertainty factors: are ratios NOAELs informative?", *Risk Analysis* 19(2):295-308.

Broome, J. 2002, "Practical Reasoning" in Bermùdez, J. and Millar, A. (2003) *Reason and Nature: Essays in the Theory of Rationality*, Oxford University Press.

Calabrese, E.J., Baldwin, L.A. and Holland C.D. 1999, "Hormesis: A Highly Generalizable and Reproducible Phenomenon With Important Implications for Risk Assessment", *Risk Analysis* 19(2):261-281.

Ditlevsen, O. 1997, "Structural reliability codes for probabilistic design – a debate paper based on elementary reliability and decision analysis concepts", *Structural Safety* 19(3):253-270.

Ditlevsen, O. and Madsen, H.O. 2004, *Structural Reliability Methods*, Internet Edition 2.2.1, http://www.mek.dtu.dk/staff/od/books.htm.

Dourson, M.L., Felter, S.P and Robinson, D. 1996, "Evolution of Science-Based Uncertainty Factors in Noncancer Risk Assessment", *Regulatory Toxicology and Pharmacology* 24(2):108-120.

Gaylor, D.W. and Kodell, R.L. 2000, "Percentiles of the product of uncertainty factors for establishing probabilistic reference doses", *Risk Analysis* 20(2):245-250.

Gaylor, D.W. and Kodell, R.L. 2002, "A Procedure for Developing Risk-Based Reference Doses", *Regulatory Toxicology and Pharmacology* 35:137-141.

Gayton, N., Mohamed, A., Sorensen, J.D., Pendola, M. and Lemaire, M. 2004, "Calibration methods for reliability-based design codes", *Structural Safety* 26(1):91-121.

Gigerenzer, G. and Todd, P.M. 1999, *Simple Heuristics That Make Us Smart*, Oxford University Press.

Herrman, J.L. and Younes, M. 1999, "Background to the ADI/TDI/PTWI", *Regulatory Toxicology and Pharmacology* 30:S109-S113.

Kalberlah, F., Föst, U. and Schneider, K. 2002, "Time Extrapolating and Interspecies Extrapolation for Locally Acting Substances in case of Limited Toxicological Data", *Annals of Occupational Hygiene* 2:175-185.

Levi, I. 1981, *The Enterprise of Knowledge*, The MIT Press.

National Research Council 1996, *Science and Judgment in Risk Assessment*, Taylor & Francis.

RIAP 1994, *Choices in Risk Assessment*, Sandia National Laboratories.

Richardson, H.S. 2004, "Satisficing: Not Good Enough" in Byron, M. (2004) *Satisficing and Maximizing*, Cambridge University Press.

Sahlin, N-E. and Persson, J. 1994, "Epistemic Risk: The Significance of Knowing What One Does Not Know" in Bremer, B. and Sahlin, N-E. 1994, *Future Risks and Risk Management*, Kluwer.

Simon, H.A. 1972, "Theories of Bounded Rationality" in Simon, H.A. 1982, *Models of Bounded Rationality: Behavioral Economics and Business Organization*, The MIT Press.

——. 1976, "From Substantive to Procedural Rationality" in Simon, H.A. 1982, *Models of Bounded Rationality: Behavioral Economics and Business Organization*, The MIT Press.
——. 1977, *The New Science of Management Decision*, Prentice-Hall.

Nanomachine : One Word for Three Different Paradigms

Bernadette Bensaude-Vincent and Xavier Guchet
Centre d'histoire et de philosophie des sciences
Université Paris

Abstract
Scientists and engineers who extensively use the term "nanomachine" are not always aware of the philosophical implications of this term. The purpose of this paper is to clarify the concept of nanomachine through a distinction between three major paradigms of machine. After a brief presentation of two well-known paradigms - Cartesian mechanistic machines and Von Neumann's complex and uncontrolled machines – we will argue that Drexler's model was mainly Cartesian. But what about the model of his critics? We propose a third model - Gilbert Simondon's notion of concrete machines – which seems more appropriate to understand nanomachines than the notion of "soft machines". Finally we review a few strategies currently used to design nanomachines, in an effort to determine which paradigm they belong to.

Key words: complexity, concretization, machine, nanotechnology, philosophy

Introduction

The convergence of nanotechnology, biotechnology, information technology and cognitive sciences, officially encouraged by the NSF under the label NBIC since 2002, has been prepared by a number of multidisciplinary collaborations. Among them, the 1997 Albany Conference on "Biomolecular motors and nanomachines", aimed at exchanging information and ideas between the research community of physicists, chemists and biologists, suggests the meeting point is the notion of machine. Five years later the convergence between nano-engineering and molecular biology materialized in the form of an electronic circuitry using a living bacterium.[1] Engineering and hybridizing inorganic and organic materials to design functional structures is now one of the most promising technological routes that will presumably produce common artefacts in the next few decades. Whatever the potential of such hybrid aretefacts, nanotechnologies and biotechnologies are presently converging is their linguistic practices. The metaphor of the machine is undoubtedly the pivot of their convergence.

On the one hand, in the biology community the machine metaphor has superseded all alternative metaphors, such as the image of the cell as a society, for instance. Cells' molecular components are described as tools or machines operating at the macromolecular level**:** Ribosomes are assembly lines, myosins are motors, polymerases are copy machines, proteases and proteosomes are bulldozers, membranes are electric fences, and so on. Although biologists generally agree that living systems are the product of evolution rather than of design, they describe them as devices designed for specific tasks. It is not that descriptions of organisms and cells as little factories are quite novel. Such metaphors were occasionally used for teaching or popularizing purposes. But following the introduction of the genetic code in the early times of molecular biology these metaphors became more than expository devices. Now the machine seems to be a heuristic model, guiding the interpretation of experiments. Even though a number of biologists confess that

[1] In 2002 NASA Argonne laboratory made circuits smaller than micro-circuits by using genetically modifying proteins extracted from high temperature tolerant bacteria as templates to create hexagonal patterns on which nanoparticles of gold were added.

the model is not to be taken literally and that the notion of program is just a cliché[2], they use the metaphor as a convenient language, providing clues about the inner functioning of living systems.

On the other hand, nanotechnology can be seen as the outcome of the new approach to nature initiated and developed by Materials Science and Engineering since the 1960s, with the core notion of "design". Materials, unlike matter, are "for something". Their structure has been processed to perform a specific task. The functional approach reconfigured the intellectual space by merging Science and Engineering. [3]It has also affected the language of chemists and materials scientists who adopted the terms "devices", "motors" and occasionally "machines" because they are concerned with the design of functional structures.[4] In looking for multi-functional and efficient materials they frequently take their inspiration from nature: spider silk, abalone shell, or lotus leaves provide engineers with model materials that they seek to mimic by their own ways and with their own tools. Some of them describe nature as an "insuperable engineer" and use such phrases as "nanosciences aim at investigating …how matter self-industrializes".[5]

The convergence of nanoscience and biology is nurtured by the shared assumption that nature works as human beings do: All its operations are supposed to be based on "devices", designed to achieve specific functions, although scientists and engineers are unable to ascribe a definite function to each part of each "natural device". It is not a trivial assumption. However, it is striking that the users of such metaphors do not care for refining their underlying assumptions and are content with a rather vague notion of machine. They use the terms "machine", "machinery", and "device", more or less interchangeably. As the machine metaphor spreads to molecules, proteins, cells … the concept looses in comprehension what it gains in extension. Since we know that linguistic practices matter, that metaphors are not neutral and have an impact on technological choices.[6]

This paper is an attempt at clarifying the notions of machine used by nanoscientists in various contexts and outlining the philosophical assumptions underlying such linguistic uses. What do nanoscientists mean by molecular motor or molecular machinery? Is it just a convenient metaphor or is it a heuristic model for understanding how nature works? And what kind of machine do they have in mind: a classical mechanical system such as Cartesian automata or something like complex systems "made up of many elements interacting in nonlinear ways", with unpredictable and spontaneous behaviors (the so-called "emergent properties")? [7] This alternative deserves particular attention because of the controversial issue at stake. Part of the concern about NBIC is related with the possibility of making molecular machines that would be out of control because of their capabilities for self-organization, self-reparation and self-replication. The latter prompted the famous grey goo scenario -– the putative result of the action of replicators breeding out of control. The relations between complexity and uncertainties about the future have been

[2] See Maurel, M.-C., Miquel, P.-A. (2001)

[3] Bensaude-Vincent (2001)

[4] Supramolecular chemists for instance used such metaphors before the term nanotechnology was coined. See for instance Jean Marie Lehn (1985). This paper has been a source of inspiration for Drexler, see Drexler, E. (1986) p. 244.

[5] See for instance Saunier, C. (2005) vol 1 p. 65. On p. 70, one can read « DNA computer tries to take inspiration from a rather efficient model of computer existing in nature, i.e. living organisms ».

[6] According to J.L . Austin's theory of speech-acts, the function of language is not only descriptive but performative. The scientific effectiveness of metaphors in biology is illustrated in Fox Keller, E. (2005). It is important to try to assess the impact of this loose terminology on the future artefacts that will be manufactured. In particular, the machine metaphor may express a deep change in the relations between nature and artefact that would consequently affect the patent policy.

[7] Dupuy, J.-P. (2000) p. 7

emphasized in particular by Jean-Pierre Dupuy. [8] He argues that by achieving complexity, converging technologists are doomed to behave as sorcerer apprentices, or at least to engage technological practices in an era of non-control. It is therefore important to closely examine what kind of nanomachines are being described and designed. Are they classical machines shrunk to the scale of atoms and molecules or are they complex systems that would gradually have the capacities to escape the control of their creators? In other words, how will nanomachines affect our relation to the material world?

After a description of the Cartesian paradigm of mechanistic machines and Von Neumann's paradigm of complex and uncontrolled machines – we will argue that Drexler's model was mainly Cartesian. In order to understand the model of his critics we propose a third model - Gilbert Simondon's notion of concrete machines. We will then review a few strategies currently used to design nanomachines in an effort to determine which paradigm they belong to.

Preliminary definitions

In this paper we take the terms nanoscience and nanotechnology in their broad and common uses, as "the study of phenomena and manipulation of materials at atomic and macromolecular scales, where properties differ significantly from those at the larger scale". [9] This definition retaining two aspects - the length scale and the emergence and exploitation of size-sensitive properties – embraces the craft of artefacts "atom by atom" or by manipulating a single molecule. It also includes certain aspects of materials science, supramolecular chemistry and bioengineering, fields that antedated the emergence of nanoscience and have extended their scopes to the nanoscale. In this broad perspective, the core project of nanotechnologies is to take advantage of the properties emerging at the scale of nanometer and to turn nanostructures into functional materials. [10] Science and technology are thus tightly interwoven. Making nanomachines and knowing how atoms and molecules behave are indistinguishable programs.

While we choose to adopt a loose notion of nanotechnology we need more precision for the notion of machine. The standard definition of nanomachine (also called nanite) as "a mechanical or electromechanical device whose dimensions are measured in nanometers" is too loose for our analysis. [11] Let us start with more refined definitions.

The term "device", coming from the French term *devis* itself forged on the Latin verb *dividire* (to divide) does not include parts[12]. It is "a thing made for a particular purpose, especially a mechanical or electronic contrivance". Like machines devices are made on purpose, to the point that it is the only idea retained in the second meaning listed in the OED "a plan, a scheme or trick". But even when a device involves various operations, there is no effort at creating a sequence generating one movement after the other.

The term "machine" coming from the Greek *mekhos*, which gave *mekhanê*, retains the connotation of trickery. It means contrivance, something ingenious and even cunning. In medieval times it was associated to forgery. According to Hugh of Saint-Victor the term *machina* derived from *moicheia* (adultery). The machine feigns to perform a natural work, like the

[8] Dupuy, J.-P. (2004).
[9] The Royal Society and the Royal Academy of Engineering, 2004.
[10] This broad meaning of nanoscience is in stark contrast with Joachim's narrow definition of nanoscience. Joachim, C. (2005)
[11] This definition was used by George Whitesides in his criticism of Drexler. Whitesides, G. (2001)
[12] Most nanoscientists do not care for the difference between a machine and a device, even though many of them emphasize that the goal is actually to let a single molecule functioning for a specific task.

adulterer feigns and pretends to be a husband. [13] Machines and alchemical operations were both considered as *hubris*, as illegitimate attempts at overtaking nature, and challenging God's creation. The current OED definition includes two meanings. The first one - " i) an apparatus using mechanical power and having several parts for performing a particular task" - emphasizes that machines have a finality, they are meant for a specific function; the latter one - " ii) an efficient and well-organized group of powerful people" - , is close to the French term "*machination*", meaning a stratagem or conspiracy. In both cases, a machine necessarily requires multiple components. In *Engines of Creation*, chapter 1, Eric Drexler quoted the definition provided by *The American Heritage Dictionary of the English Language*: "Any system, usually of rigid bodies, formed and connected to alter, transmit, and direct applied forces in a predetermined manner to accomplish a specific objective, such as the performance of useful work."[14] Three aspects are noticeable in this definition: i) a machine is made on purpose out of rigid or stable components ; ii) a machine is something which converts energy and transfers forces in a specific direction ; iii) a machine is meant to produce work, to perform useful tasks. All machines whether they be simple machines like levers or combustion engines or information machines fulfil at least the three requirements. Nanomachines will have to do the same if they pretend to be machines.

Cartesian and complex machines

Within this notion of machine two paradigms have been distinguished. The earlier paradigm, which stabilized in the seventeenth century, was modelled on the mechanical automata described by Descartes and materialized by artists such as Jacques de Vaucanson, among others. The more recent one is the paradigm of complexity, supported in particular by John Von Neumann at the Hixon Symposium (Caltech) in September 1948.

A Cartesian automaton - such as pumps, gears, levers- is a multi-component machine designed to produce movements. It can be divided up into parts, like difficulties in Descartes's first rule of the *Discourse of method*. To its designer, a Cartesian machine is transparent, perfectly understandable and predictable, without mystery. The designer (clockmaker or engineer) has a full control over his machine because he has designed each component and their details. The Cartesian machine is *partes extra partes*, each part being independent has to be assembled to the others (wired, clipped or welded). Each individual component is ascribed a definite function, which is its *raison d'être*. The parts are independent but they have no individuality. They contribute to the whole but the whole does not maintain them.[15] Each part is necessary, none is sufficient. Each one is made on purpose to fit into the system and has to be adjusted to the others. A machine is exquisitely functionalized in all its details. As the French philosopher George Canguilhem argued a machine is much more teleological than living organisms.[16]

Von Neumann's General and Logical Theory of Automata[17] was developed as an alternative to the model of the central nervous system shaped by the cybernetician Warren McCulloch and Pitts. They had described the brain as a computing machine, a communication network of

[13] Jerome Taylor ed. *The Didascalion of Hugh of Saint Victor* (New York, Columbia University Press, 1981, pp. 55-56, quoted from Newman, W. (1989), p. 424.
[14] *The American Heritage Dictionary of the English Language*, edited by William Morris, Boston, Hougthon Mifflin, 1978, in Drexler, E. (1986) p. 5.
[15] See Canguilhem, G. (1979), « La partie et le tout dans la pensée biologique » and his distinction between the technological model and the political model (le tout est aussi au service des parties, l'organisme entier contribue à la vie des cellules)
[16] Canguilhem, G. (1952)
[17] Neumann, J. Von (1951)

elementary arithmetical calculators (neurons) that compute a function of their antecedents. This machine would work provided neurons be activated by stimuli beyond a critical point. Von Neumann emphasized that it was still possible to describe the behavior of McCulloch's logical machine in a finite number of words. The structure of the machine was much more complicated than the model describing its behavior. But what about automata who have a behavior so complicated that it is impossible to characterize it fully in a finite number of words? In that case, Von Neumann argued, it would be simpler to describe its structure. The best model of the automaton would be the automaton itself. This is a complex machine. Instead of designing a structure to perform a predefined task (the function determining the structure), you have to build the structure in order to know what is capable of.

The contrast between Cartesian machines and complex machines also concerns the part/whole relationships. Cartesian machines are artificial totalities, i.e., the parts exist prior to the whole and the whole is nothing but the sum of its components. Cartesian machines, just as Mc Culloch's computing machine, are devices transforming inputs into outputs. By contrast, complex machines are close to natural totalities. Unlike aggregates whose unity is accidental, they are made up of various elements interacting in loosely determined ways, and resulting in non-linear effects. Complex automata are autonomous, self-organized totalities made up of several integrated levels with a hierarchy of structures. From the interaction between the elements, a spontaneous and collective order emerges. The properties of the machine are novel and non-deducible from the properties of the elements. In return, the emergent order imposes constraints on elementary interactions. "The whole and its elements therefore mutually determine each other" [18] This codetermination relies on feedback loops between the various levels, and specifies the notion of complexity in artificial and natural automata.[19]

Finality makes a third major difference between Cartesian and complex machines. A Cartesian machine is heteronomous, as the purpose is not the machine itself. The intention is part of the definition of the machine: such a machine is designed *to* perform a defined task. The machine pre-exists in the mind of the designer. It thus instantiates the subjective notion of finality: the designer's intentions are embedded in the mechanism, which is just their materialization. A perfect machine will be the one presenting a strict isomorphism between the subjective goal and the objective mechanism.

By contrast, a complex machine is autonomous in the sense that it is not translating any subjective goal. The major feature of a complex machine is that it escapes from the control of its inventor. Its behavior is strictly unpredictable, so that one has to wait and see the machine in action in order to know how it behaves. As Dupuy often emphasizes, in complex machines the designer's purposes have to be superseded by the machine. The fear of the sorcerer's apprentice subdued by his own creation is not a potential hazard, an accident. It is the very essence of

[18] Dupuy, J.-P (2000)

[19] Von Neumann's talk was not the sole attempt at the Hixon Symposium to introduce the notion of complexity against McCulloch's constructive approach. The neurophysiologist Karl Lashley, and the embryologist Paul Weiss also argued that the brain was not a computing machine, and rather was a continuous field with emergent features. Although Lashley and Weiss's approach to the nervous system was clearly antireductionnist (irreducible to their components), it was not holistic: complex totalities are neither reducible to the properties of their parts, nor Leibnizian monads whose unity is substantial. Between reductionism and holism, between nominalism and substantialism, the theory of complexity offered a third model of the whole/parts relations. Unlike Von Neumann however, Lashley and Weiss drew a sharp boundary between living and non-living beings. For them, complexity was the exclusive property of biological systems whereas Von Neumann assumed that complexity could be embedded in artificial automata.

complex machines. Von Neumann himself prophesized that "the builders of automata would find themselves as helpless before their creations as we ourselves feel in the presence of complex natural phenomena".[20]

According to Dupuy, this sort of machine is what nanoengineers have in mind. The lack of control is an essential feature of nanotechnology, although it is not necessarily linked to the existence of self-replicating devices such as Drexlerian replicators.

"In keeping with that philosophy the engineers of the future will not be any more the ones who devise and design a structure capable of fulfilling a function that has been assigned to them. The engineers of the future will be the ones who know they are successful when they are surprised by their own creations […].It will be an inevitable temptation, not to say a task or a duty, for the nanotechnologists of the future to set off processes upon which they have no control". [21]

Most scientists and engineers active in the field of nanotechnologies are willing to demarcate their projects from what they view as speculations and fantasies. It is important however to examine if the design of complex machines is part of their program. With the conceptual distinction between Cartesian and complex machines in mind, we can now review the current literature on nanotechnology to see if there are candidates for the latter category.

Drexler's molecular manufacture

Drexler is an obvious candidate. As early as 1986, his prophecies of "molecular manufacture" were guided by the description of proteins and ribosomes in terms of machinery, and as a post-graduate, he studied in the laboratory of Marvin Minski, a leading figure of Artificial Intelligence[22]. According to Otavio Bueno, Drexler's views of self-replicating nanorobots were inspired by Von Neumann. [23] His argument is based on the evidence of a few references to Von Neumann in *Engines of Creation* and on an interview with Drexler. However the influence of Von Neumann on Drexler is far from obvious.

Drexler started with a conventional definition of machine in Chapter 1,[24] and he often claimed that his molecular manufacture was the extrapolation of today's automated factories to the smallest scale, by a process of 'mental shrinking'. "Just as ordinary tools can build ordinary machines from parts, so molecular tools will bond molecules together to make tiny gears, motors, levers [...] and assemble them to make complex machines". [25] He described molecules as rigid building blocks, similar to the parts of tinker toys to be assembled like the elements of Lego construction sets. The functions performed by the various parts of molecular machinery are also essentially mechanical. They position, move, transmit forces, carry, hold, store, etc. The assembly process itself is described as a "mechanosynthesis", positioning the components with a mechanical control.

However there are four occurences of the phrase "complex machines" in *Engines of Creation*. One relates to protein machines: "the forces that stick proteins together to form complex machines are the same ones that fold the protein chains in the first place". [26] The others are related to artificial machines: "Just as ordinary tools can build ordinary machines from parts, so

[20] Dupuy, J.-P. (2000) p. 142
[21] Dupuy, J.-P., Grinbaum, A. (2004) p. 8
[22] Drexler got his PhD laboratory at MIT in Marvin Minski's, who in turn had been supervised as a doctoral student by Von Neuman. Minski wrote a preface for *Engines of creation* in 1986
[23] Bueno, O. (2005).
[24] Drexler (1986) p. 5
[25] Drexler, E. (1986) p. 12. See also Drexler, E. (2001), p. 74.
[26] Drexler, E. (1986) p. 10.

molecular tools will bond molecules together to make tiny gears, motors, levers [...] and assemble them to make complex machines" [27]; the third one concerns the feasibility of nanotechnology and assemblers, "the heart of the case rests on two well-established facts of science and engineering. These are (1) that existing molecular machines serve a range of basic functions, and (2) that parts serving these basic functions can be combined to build complex machines". [28] Finally, each advanced assembler can contain "an average of one hundred atoms – enough parts to make up a rather complex machine" .[29]

The three references to artificial complex machines derive from bioengineering, which globally rests on the view of cells as factories full of individual machines. In Drexler's view, genetic engineers have full control on the individual machines. They pick and place them, they re-engineer DNA and proteins in order to perform pre-determined specific tasks. In short, they rely on a Cartesian paradigm. Although he never refers to Descartes, Drexler shares his famous claim that the combinations of the visible parts of our machines are analogous to the combinations of the tiny (of course Descartes didn't say "nano") invisible components of animal organisms.[30] "Molecules have simple moving parts, and many act like familiar types of machinery". [31]

Drexler nevertheless stressed a big difference between cells and artificial machines. Unlike our machines, natural molecular "machines" (in cells) are self-assembling. If we put the different parts of a car in a big box, and if we shake the whole, we never get a car. Drexler's program comes down to reduce this ultimate difference. Unlike bulk technology, molecular technology allows a way for parts to self-assemble. Tomorrow's nanoengineers will design artificial nanomachines, new protein tools that will be able to assemble parts. They will act like automated machine tools programmed by punched tapes. These programmable protein machines inspired by ribosomes and enzymes, will bond molecules together with great precision. They will be made of a tougher stuff than the soft and weak molecular machines of the cell.

"Protein machines will thus combine the splitting and joining abilities of enzymes with the programmability of ribosomes […] Enzyme-like second-generation machines wil be able to use as "tools" almost any of the reactive molecules used by chemists – but they will wield them with the precision of programmed machines". [32]

Drexler's programmed assemblers have nothing in common with Von Neumann's automata. The universal assembler is not self-replicating. It needs material and energy, and instructions for use. His molecular manufacture made *partes extra partes*, with assembling process, is a mixture of conventional mechanics and computer science. A complex machine in Drexler's view is just an aggregate of simple machines. Insofar as he relies on the view of both natural and artificial machines as systems reducible to their parts, Drexler has no choice but to describe the assembly process by analogy with a macro manufacture.

Descartes's analogy between living beings and artificial machines presupposed the fiction of an artisan-God manufacturing natural bodies parts after parts. Indeed Drexler does not explicitly need such metaphysical requisit, although his nano-fingers have the creative power of God's finger. Drexler's world is in the hand of a magic engineer, the so-called "replicator", which inspired the grey goo scenario based on a process of uncontrolled self-replication. A replicator is made of a reader, a tape, several assemblers and other nanomachines. According to Richard

[27] Drexler, E. (1986) p. 12. See also Drexler, E. (2001), p. 74.
[28] Drexler, E. (1986) p. 17.
[29] Drexler, E. (1986) p. 56.
[30] Descartes (1637), *Discours de la méthode*, 5th section.
[31] Drexler, E. (1986) p. 102
[32] Drexler, E. (1986) p. 14

Dawkins (quoted by Drexler), a replicator is a thing that makes a copy of itself. RNA molecules and cells qualify. Replicators manufacture nanosystems by means of assemblers, such as cells manufacture proteins by means of ribosomes, and they are supposed to bridge the gap between human and natural machines. [33] Drexler suggests a sort of "network of factories" forming a self-expanding, self-replicating system. In such a system, "robots could do all the robots-assembly work, assemble other equipment, make the needed parts, run the mines and generators that supply the various factories with materials and power, and so forth". [34]

Here automated engineering and molecular manufacturing are closely intertwined. But could we go further and characterize replicators as complex machines in the sense of Von Neumann? Replicators have two remarkable features of complex machines: autonomy and self-replication. Drexler remained elusive on the feasibility of his replicators. He just mentioned that: "the chief requirement will be programming the first replicator, but AI systems will help with that. The greatest problem will be deciding what we want". [35] It comes to no surprise that the controversy raised by Drexler focused on the feasibility of his self-replicating nanorobots. As Whitesides argued: "The assembler, with is pick-and-place pincers, eliminates the many difficulties of fabricating nanomachines and of self-replication by ignoring them". It is clear that Drexler did not really explore the feasibility of such complex machines. In fact, Drexler confessed that his concept of molecular manufacture does not require self-replicating nanorobots, when confronted to the public anxieties raised by this fiction, he admitted "I wish I had never used the term 'grey goo'". [36] The fact that he could so easily drop his replicators, suggests that they were just one more independent piece of his machinery, performing a specific task. They were parts of a Cartesian machine.

To sum up, Drexler's molecular manufacture is described as a collection of independent parts even in its effort to include attributes of complex machines. His grand vision basically rests on a mechanical view of machines combined with the literary theme of the uncontrolled robot. The choice of the term "robot" coined by Karel Capek in the context of utopian (or dystopian) literature, is an indication that his work belongs to the literary genre of science fiction rather than to technical literature on automata. The image of the grey goo revitalized a literary tradition expressing the public's fear of technology.[37]

Drexler's model has been submitted to merciless critics by chemists such as Richard Smalley and George Whitesides, and other scientists who clearly established that Drexler's model of machine was inadequate to operate at the nanolevel.[38]

Soft Machines or Concrete Machines

[33] Drexler, E. (1986) p. 56: "Some of these replicators will not resemble cells at all, but will instead resemble factories shrunk to cellular size. They will contain nanomachines mounted on a molecular framework and conveyor belts to move parts from machine to machine. Outside they will have a set of assembler arms for building replicas of themselves, an atom or a section at a time".
[34] Drexler, E. (1986) p. 54
[35] Drexler, E. (1986) p. 121
[36] Phoenix, C., Drexler, E. (2004)
[37] Daniel P Thurs and Stephen Hilgartner rightly noted that the threat of the expansion of the grey goo is the mirror image of the threat of an uncontrolled public opinion – like the luddites or the opponents to GMOs refusing new technologies. See Conference on nanoethics, South Carolina, March 2005.
[38] See articles by Richard Smalley, George Whistesides, Robert Buderi in *Scientific American,* Sept 2001. Chris Phoenix, "Of chemistry, Nanobots and Policy", Center for Responsible nanotechnology, December 2003.

Drexler's machines have been proved non feasible because they are not adapted to the special features of the nanoworld. As Whitesides emphasized a nanoscale submarine would be impracticable because of Brownian motion, which would make useless all efforts to guide the submarine. However neither Smalley nor Whitesides did try to promote an alternative concept of machine.[39]

Philip Ball pointed to chemistry as an alternative to the mechanical approach:

> I feel that the literal down-sizing of mechanical engineering popularized by nanotechnologists such as Eric Drexler - whereby every nanoscale device is fabricated from hard moving parts, cogs, bearings, pistons and camshafts - fails to acknowledge that there may be better, more inventive ways of engineering at this scale, ways that take advantage of the opportunities that chemistry and intermolecular interactions offer.[40]

Richard Jones, another critic of Drexler's machines tried to delineate the profile of more plausible nanomachines. His concept of "soft machines" was a clear response to Drexler rigid machines and mechano-synthesis. Whereas Drexler's assemblers were downsized versions of familiar machines, Jones stresses that nanomachines cannot be small-scaled versions of industrial macromachines, because of the special physics of the nanoworld. "Physics is different in the nanoworld, and the design principles that serve us so well in the macroscopic world will lead us badly astray when we try to apply them at these smaller scales".[41] It means that engineers will have to abandon their familiar frameworks. Jones encourages a decisive step : to start addressing the question « how artefacts will function » prior to "how are they to be made », [42] His conviction is that the model for nanoengineering lies in biology. Jones argues that biological soft machines are not the outcome of "the unhappy consequences of the contingencies of evolution", rather they are "the most effective way of engineering in the unfamiliar environment of the very small".[43] In his view, biological mechanisms and materials have been designed at the nanoscale, they are perfect to work at that level, they are completely adapted to the special physics of the nanoworld, even though they are not always efficient at the macroscale.

> A steam engine is better than a horse, strong and lightweight aluminium alloy is a better material to make a wing out of than feather and bone […] Big organisms like us consist of mechanisms and materials that have been developed and optimised for the nanoworld, that evolution has had to do the best it can with to make work the macroworld".[44]

Biology would be then the unique model for engineering at the nanoscale. Therefore Jones outlined the general principles of biological molecular processes and pointed out three major differences between the bio-machines and human conventional technologies. a) Instead of

[39] When Whitesides asked "What is a machine?", he contented himself with a very traditional answer. "A machine is a device for performing a task". It has "a design, it is constructed following some process, it uses power, it operates according to information built into it when it is fabricated". [Whitesides, 2001, p. 78]
[40] Ball, p. (2002), p. 16
[41] Jones, R. (2004), p. 85
[42] See Jones' Softmachine Blog : entry Wednesday, June 29th, 2005 « Debating the fesaibility of nano manufacturing »
[43] Jones, R. (2004), p. 2, 3
[44] Jones, R. (2004), p. 6, 7

channelling the traffic with tubes and pipes, living systems take advantage of Brownian motion, which moves molecules around and continuously bombard nano-objects. b) Living systems do not use rigid molecules like synthetic chemists: molecules easily change shape and conformation. c) The constraints in building machines at the molecular level differ from those of "bulk technology". Inertia is no longer a crucial parameter but surface forces – viscosity- becomes a major constraint that prompts nano-objects to stick together.[45] Whereas Drexler considered the distinctive features of the nanoworld as obstacles to be overcome by means of tricks, Jones insists that nanomachines will have to do with Browian motion. Nanomachines will not be designed until engineers abandon their "conventional engineering" and invent new concepts of machines. The key is to understand that "a different feature of the physics that leads to problems for one type of design may be turned to advantage in a design that is properly optimised for this different world".[46] The properties characteristic of the nanoscale, which are problems for conventional machines, will have to be used as positive opportunities by nanoengineers. Jones thus contrasted two "design philosophies" to make nanoscale artefacts. Conventional design is based "on the principles that have served us so well on the macroscopic scale would rely on rigid materials, components that are fabricated to precise tolerances, and the mutually free motion of parts with respect to each other. As we attempt to make smaller and smaller mechanisms, the special physics of the nanoworld - the constant shaking of Brownian motion and the universal stickiness that arises from the strength of surface forces - will present larger an larger obstacles that we will have to design around". [47] Nanodesign should be based on the principles used by cell biology, labelled 'soft engineering'. "The advantage of soft engineering is that it does not treat the special features of the nanoworld as problems to be overcome, instead it exploits them and indeed relies on them to work at all".[48]

Changing obstacles into positive principles of work is exactly what the French philosopher Gilbert Simondon called "concretization". In his famous book *Du mode d'existence des objets techniques* (1958), Simondon elaborated a new concept of machine, which differed both from the cartesian model of mechanistic machines and from Von Neumann's concept of complex machines. He started with a general distinction between abstraction and concretization. A machine is "abstract" when each part has been designed independently, each one for a definite and unique function. Cartesian machines are typical abstract machines because the concept of the machine in the designer's mind precedes the machine itself. The operations performed by the machine result from its conceptual consistence: there is nothing more in the machine than in the designer's mind. And of course the machine has to be built before it starts to operate.

By contrast, a concrete machine would not be deduced from general principles. Its feasibility depends on its operating conditions rather than on scientific principles. In fact, it is the machine itself, which creates the conditions required for its operation. The environment where the machine will operate is not an external feature or a simple parameter that engineers have to take into account in the design process. The milieu is not something to which the machine will have to be adapted; it is an intrinsic aspect of the design of the machine. A concrete machine works precisely because of (and not despite) its association with a specific environment.

Simondon illustrated the contrast between abstract and concret machines with the example of a hydraulic power station, known as Guimbal's turbine. The problem was to build an electric generator, small enough to be immersed into a water pipe. The major obstacle was the heat

[45] Jones, R.A. (2004), p. 56-86.
[46] Jones, R. (2004), p. 86
[47] Jones, R. (2004), p. 127
[48] Jones, R. (2004), p. 127

produced by the generator, which would cause its explosion at a critical point. Conventional engineers would typically look for all physics principles in order to reduce the size of the generator and subsequently prevent its explosion; then they would adapt the system for underwater conditions. The machine resulting from this conventional design is what Simondon labelled an "abstract" machine. By contrast the "concrete engineer" will imagine how an immersed generator would work, instead of striving to make the generator smaller and smaller before introducing it in a water pipe. The generator has to be in a container filled up with oil. It is supposed to be coupled to the turbine by means of an axe, and immersed into the pipe. In this configuration, water will perform various functions : it supplies power to the turbine, it keeps the machine working; it also exhausts the heat generated by the rotation of the turbine. Oil is also multifunctional: it lubricates the generator; it conveys the heat released by the generator to the surface of the container, which is cooled by the water; and it prevents water to come into the container, due to the difference of pressure between oil and water. The two liquors thus cooperate : the faster the turbine and the generator are rotating, the greater the agitation of oil and water will be, and the better is the cooling of the system. As Simondon emphasized, the aqueous milieu determined the design of the generator. The Guimbal turbine would never work in open air: it would explode. The concrete machine is tightly associated with its environment (in this case, the couple oil and water). Simondon calls *individu technique* (a technological individual) such a machine because it is self-conditioned, it does not exist as a possible machine prior to being in operation. Since the interactions between the various elements of the machine are not deducible from any set of scientific laws, technology is not science-based. It follows that there is always more in a working machine than in the mind of its inventor.

At this point Simondon introduced a second distinction between "constitution" and "invention". In his view, the constitution of artefacts is just the materialization of an abstract machine. All effects can be deduced from the analysis of the concept of the machine. Design and operation are two independent tasks. By contrast, to "invent" a machine, is not just assembling logical functions and then put the system in action. The machine is designed according to its operating conditions and in fact, it invents its own environment. The associated environment cannot be anticipated and becomes integral part of the machine. Therefore the "mode of existence" of a "technological individual" cannot be defined prior to its functioning. [49]

Simondon's concrete machines thus deeply differ both from Cartesian machines and from programmed automata. They are not built *partes extra partes* but invented straight off by envisioning, "imaging" the feedback loops between the machine and its *milieu associé*. But do they also differ from Von Neumann's complex machines? To a certain extent, the system made up by a concrete machine and its associated environment is complex. First, since the machine is self-conditioned, it is autonomous and Simondon suggested that concrete machines were close to the mode of existence of natural beings and that engineers should deal with them as they do with living beings. Second, concrete machines are unpredictable since their inventors will not know how to make the machines until they actually start building them. However, unlike Von Neumann's complex automata, Simondon's concrete machines are not self-replicating and their unpredictability does not mean that they are out of control. Never did Simondon suggest that we were about to face a terrifying lack of control over human artefacts. On the contrary, the incorporation of special features of the associated environment into the machine, and the conversion of external data into essential working conditions (such as oil and water in the example of Guimbal's turbine) warrant a better control on the system. Indeed the machine supersedes the plan that its inventor had in mind, but it never supersedes the inventor. More precisely, by contrast with Von Neumann's approach to complexity, a concrete machine still

[49] Simondon, G. (1989)

relies on the reference to a human subject. Such a machine involves the very special ability of human beings to stress analogies between biological and technological operations. Simondon assumed that we can invent self-conditioned machines because we are ourselves self-conditioned living beings. To be sure, Simondon's subject is no longer a Cartesian *maître et possesseur de la nature*. Nevertheless concrete machines rely on human subjects.

To sum up this section, the strong similarity between Simondon's concrete machines and Jones's soft machines rests on two key ideas: looking first at how the machine will function and turning obstacles into conditions. However thanks to its additional features - individuality, incorporation of the milieu, and reference to a human subject - Simondon's notion of concrete machine may provide us with more robust conceptual resources for understanding what is going in nanotechnology, than Jones's metaphorical notion of soft machines.

Now that the controversy raised by Drexler seems to be closed, and Drexler marginalized, it is time to examine what kind of nanomachines are being effectively designed in laboratories and (maybe for the near future) in manufactures. Are nanoscientists and engineers designing conventional Cartesian machines, or are they aiming at creating uncontrolled machines in the sense of Von Neumann and Dupuy, or something more akin to Simondon's concrete machines? Let us look at a sample of machines described in scientitifc publications. Of course the purpose of this review is not to make a kind of "philosophical evaluation". It is rather aimed at encouraging reflections on the ways of designing nanomachines.

Nanorobotics and Smart Structures

In September 2004 many newspapers reported a "mechanical miracle". Metin Sitti, director of the Nanorobotics Lab at Carnegie Mellon University built a tiny robot that walks on water like water spiders. This artificial insect was inspired by the mode of locomotion of the *Gerridae*, a variety of water striders recently studied by an MIT team, which move at 1m/s, the equivalent of 700km/h. Sitti's prototype raised great excitement because it could be equipped with chemical sensor to detect contaminants in water or with a camera to act as a spy. But what kind of machine is it? The body is made of carbon fibres linked to eight steel-wire legs coated with water repelling plastic. Its "muscles" are flat plates of piezoelectric material. The power is supplied and controlled through three circuits. The "miracle" is precisely that it is a simple automaton. As Setti emphasized those insects have no brain, they don't need brain with such simple control. [50] Indeed it is a tiny insect–1 gram – but it is not nano, at all. Using only piezoelectricity (the property of changing shape under pressure to produce electricity) for the actuator, it does not rely on size-dependent properties.

Building up true nanorobots confronts us with a communication problem. How to exchange instructions, energy or information with nano-scale objects? Their manipulation with macroscopic instrument such as the STM is just a primitive stage. More refined tools have to be invented in order to « translate » information in quantum physical terms understandable by a nanoscale objects. This is undoubtedly a major challenge for narobotics. Yet it will lead neither to concrete nor to complex machines.

The basic principles of such robots are borrowed from Automated Engineering. They consist of a sensor, a processor and an actuator. The functions being more or less similar to those of humans these items are named ''smart" or "intelligent structures''. They are so interesting for technological applications that they have been one of the major goals of materials science over the past decade. However, these robots do not require complexity. Smart structures of Micro-Electro-Mechanical Systems (MEMS) are like Cartesian machines. One material acts as a sensor;

[50] http://www.me.cmu.edu/faculty1/sitti/nano/index.html. *Le Monde*, mercredi 15 septembre 2004, p. 25

another one as an actuator; and a third one—generally silicon—is the processor. Access to the nanoscale would increase the performances of microsensors since they could exploit the huge surface of nano-objects in order to detect biochemicals or contaminants. Ideally a nanorobot should be made of one molecule playing the role of a sensor, the next a processor, and a third an actuator. Such an ideal robot would nonetheless still be designed like a Cartesian *partes extra partes* machine with a component for each specific task and would have none of the features of complex machines or concrete machines.

Molecular motors

Molecular machines are extremely fashionable. Following the design of a variety of tools - gears, rotors, levers, tweezes, switches – in the 1990s, the design of motors has been a major concern since 2000. In fact, prior to the take off of nanoscience, a few molecules capable of moving and rotating had been designed by supramolecular chemists. For instance, the rotaxanes designed by Jean Pierre Sauvage as early as 1983 with a macrocyclic ring trapped onto a "thread" by two bulky "stoppers", were initially considered as curiosities resulting from a difficult and low-yielding synthesis. The chemists who rest on the principles of chemical topology to interlock those molecules used to describe them as "architectures" rather than as machines. Over the past decade, the few exotic molecules became a whole collection of molecular machines whose synthesis has been made easier thanks to the use of non-covalent (hydrogen bonds or metal-ligand bonds) interactions, with the help of templates to hold the molecular precursors in the correct orientation.[51]

Another example - the molecular wheelbarrow - will help to "anatomize" a molecular motor designed from bottom-up. The designers of the molecular wheelbarrow use the phrase "technomimetic molecules", since their project was to recreate at the molecular level the functions of macroscale machines. Interestingly they define the molecular machine as a "molecule responding to the orders of its operator". Whether the operator is the tip of a STM, another molecule or a human hand, the concept is the same. The molecular machine is under control and it has no autonomy whatsoever. The purpose of such challenges is less to make useful technological artefacts than to understand the properties of isolated molecules. After a first attempt at designing a non-directional rotor made of decacyclene in 1998, [52] Christian Joachim and his group reported the design of a uni-directional rotor. It uses a C60 molecule bouncing between two electrodes to transport individual electrons from the source to the drain. The dissymmetric position of the molecule allows the control of the rotation movement. The wheelbarrow consists of the rotor (C60), a stator and a ball-joint (ruthenium atom). Its body itself is an organometallic structure shaped as a three-leg piano stool. The wheelbarrow does not move as its designers predicted. And the identification of the obstacles is probably the most interesting result that they could get. One reason is the molecule flexibility. Instead of standing rigid like crystals, it changes "like Dali's famous clocks". A second and major obstacle comes from quantum fluctuations that prevent the stabilization of the device. There is no way to control such fluctuations. Molecular designers have to make with it. Here may be is a promising pathway to generate a concrete machine capable of taking advantage of contingent fluctuations to achieve a specific task assigned to the machine by the designer.

Molecular Electronics

[51] As an example of the use of hydrogen bonds see : Leigh, D. A., Wong, J. K. Y., Dehez, F. Zerbetto, F. (2003). For an overview of molecular motors see ?

[52] Joachim, C. and al, *Science* 281 (1998) 531-33.

Up to this point we have only considered machines performing mechanical functions. What about machines performing logical tasks such as storing information, or even computing? Would molecular electronics be a more serious candidate for concretization?

Embedding computing capacities in a single molecule has been a dream since the dawn of computer age. In 1974, Mark Ratner (New York University at that time, now Northwestern) and Ari Aviram of IBM envisaged building computers from bottom-up by turning individual molecules into circuit components. This remained a thought experiment (and a stimulating dream) until the 1980s when the scanning-tunnelling microscope (STM) came into use.[53] Over the past decades a host of molecular electronic devices have been designed. And the breakthrough of 2001 was connecting those devices to make a circuitry. Indeed the step from the device level to the circuit level was a major achievement legitimizing the term nanomachine. However we are still far from both complex and concrete machines. The nanocircuit is nothing more than a collection of independent parts, each one performing a particular task. It is an "abstract" machine meant for an external purpose. There is no indeterminacy apart from the conventional margins of failure. To achieve a real move towards a non-Cartesian machine, one would have to get rid of the concept of circuitry and to design a radically new concept of electronic machine. Such a problem was clearly formulated by Christian Joachim:

> The machine that we are trying to design has no parts. Our aim is precisely to get rid of parts, be them electronic devices or Q-bits. Mechanics is still practiced in a sensorial space with parts to assemble. Such was also the case in the early times of molecular electronics. We had to divide the circuit into small parts: molecules, quantum bits. But it turned out that it is difficult to control the whole system on a wafer. Now, we are exploring a partless approach. In quantum dynamics, we deal with the space of states and no longer within the usual space. The approach is formally similar to that in thermodynamics of computation. We need to be out of equilibrium, at the quantum level by preparing the molecule in a non-stationary quantum state. The molecule has to be out of equilibrium in order to have it performing a task. But it is costly in design because we have to maintain the quantum evolution out of decoherence during one computation cycle. It is also costly in control because we have to control the full quantum trajectory in a gigantesque state space for each logic function.[54]

This project points to a new sort of machine. Will it be a complex or a concrete machine? The answer would be premature.

Wet Technology

Over the past decades molecular biologists and biophysicists have jointly investigated the motors that move muscles, sperm and cells, in living systems for a variety of medical applications. These natural phenomena are invariably described by analogy with human technology. [55] The conditions

[53] For a historical sketch of molecular electronics see Joachim, C., Gimzewski, G., Aviram, A. (2000)
[54] Personal interview, Toulouse, February 15, 2005.
[55] The “power station” fuelling “living motors” is the ATP synthase. It provides the chemical energy that proteins transform into mechanical energy for cellular locomotion, division, maintenance and intracellular

for proteins such as myosins, kinesins and dyneins to be motors have been studied for many decades, but now biologists and nano-engineers want to know how exactly they operate at the molecular level. In this respect, the research field now established as bionanotechnology differs from the research tradition in biomechanics initiated by D'Arcy Thompson. The structures and processes displayed in biology came to epitomize a new technological paradigm often labelled "wet technology" since operations in living systems are usually performed at room temperature, in aqueous milieu with soft materials much more flexible and versatile than the parts of our rigid machines.

The Bioengineering Nanotechnology Initiative launched in 2002 by the US National Science Foundation prompted a reorganization of research with interdisciplinary teams aiming at identifying the molecular components of living systems, and understanding the process of their synthesis *in situ* in order to take inspiration from them. Understanding the ways of nature and exploring new technological avenues merge into one single research program. In this program, it is more or less tacitly assumed that understanding one biological motor comes down to understanding a fundamental process because nature tends to use and re-use the same solution for a problem. And it is more or less expected that the access to the "fundamental" level secured by molecular biology will provide us with THE bottom-up method that nature and art can share. Nanotechnology and molecular biology rest on the same epistemological credo that a detailed knowledge of structure will lead to the control of functions and sometimes even processes. [56] As long as such programs tend to capture an essential structural element and rely on it while neglecting all the messiness created by molecular agitation at the nanoscale, they are not really leading to a new technological paradigm. Whatever the promise and prowes of the sophisticated nanomedicines under study, from a philosophical perspective they look extremely conventional.

At the cross-road between biology and nanotechnology, two different strategies are being used: either re-engineering biological machines for making artefacts or mimicking them, making artefacts inspired by technical solutions diplayed in nature.

Since the mechanisms designed by living systems are the best candidates for the title of complex machines, it is tempting and promising to take hold of them and divert them for technological purposes. But are we sure that re-engineering machines designed by living systems in order to perform tasks they are not meant to perform, will help build complex machines?

Molecular recognition is the most enviable property that engineers seek to use for the design and synthesis of all kinds of machines. DNA is a very efficient tool for building nanomachines, provided it be re-engineered for technological purposes. For instance, branched DNA molecules – instead of linear sequences – with sticky ends can be used as scaffolding to organize the components of nanoelectronics. DNA can also be used to produce mechanical devices because it is robust. But the huge organizational potentialities of DNA cannot be efficiently exploited unless DNA is combined with inorganic components such as nanotubes or nanocrystals whose physical properties are directly needed for applications. The "soft machines" designed by nature are not directly fit for the conditions of dry technology. Researchers have begun to harness biological structures to optimize existing functions of nucleic acids and proteins or to create new ones. As

---

traffic.

[56] This shared assumption is noticed by the anonymous editoralist of « why small matters », *Nature Biotechnology*, 21, number 10 (October 2003) p. 1113. The research program conducted by the Curie Institute in Paris on Myosins aiming at unveiling their atomic structure with the help of X-Ray Cristallography exemplifies the assumption from structure to functions.

Ronald Breaker argued, "the challenge for biochemists is to take RNA and DNA beyond their proven use as polymers that form a double helix". [57]

Although this option is sometimes considered the most promising for commercial applications, [58] from a technological perspective it may be deceiving. First nanobioengineers tend to isolate a few interesting mechanisms from their context of operations and they overlook the difference between the contexts of human design and nature's design. The former relies on plans and aims at standardization - while evolution is a blind process generating variability through mutation and recombination over a long period of time and later selecting a few structures. As Steven Vogel emphasized, each domain has acquired a coherence and consistency, a rationality of its own, so that it maybe a nonsense to pick up a few local recipes and try to copy them. [59] Moreover, the current examples of hybrid devices relying on the convergence of technologies are just designed by aggregation of functions. They are deduced from scientific principles and built up *partes extra partes*. Hybridization comes down to downplay the complex machines "invented" by nature in order to turn them into simple Cartesian machines. Hybrid machines are "constituted" rather than "invented". Even the grandiose programme aimed at making hybrid machines or robots assisting, repairing human bodies and brain, through the convergence of nanotechnology, biotechnology and cognitive science, belongs to the old Cartesian paradigm, since the basic assumption is that living organisms are "chemical computers" i.e. machines with internally stored information.[60] The brain itself is described as a machine ruled by algorithms.[61] The "mechanization of the mind" may well lead to building useful devices but less plausibly to complex machines or concrete machines.

The alternative strategy - biomimetism - has been first developed by materials scientists who realized that nature had built multifonctional and highly performant structures and that could well draw lessions from nature. This approach resulted in the design of a number of already commercialized structures as well as to better understanding of biomineralization in marine organisms or the production of fiber by spiders. However this approach does not exclusively belong to nanotechnology, since it is based on the clear recognition that the performances of natural structures are due to their hierarchical structure, and consequently involve multiple length scales.

The interest of chemists for processes as well as structures has prompted their attention – and admiration - for the process used by cells to reproduce when they divide. "Self-assembly is a process in which molecules or parts of molecules spontaneously form ordered aggregates, usually by non covalent interactions". [62] Self-assembly involves two major features. First, it is a spontaneous process. Components of living systems assemble without intervention of orders coming from outside. Instructions for the design of the "machine"are built in the components, and the environment is involved as a component. Second self-assembly uses reversible interactions, i.e. non-covalent bonds. The continuous thermal agitation allows molecules to move around, in order to adjust and re-adjust. These reversible arrangements are crucial to obtain aggregates without defect.

Self-assembly is more similar to self-organization than to conventional engineering. Creating order out of disordered moving elements is so typical of life that it has long been ascribed to a

[57] Breaker, R. (2004)
[58] For instance Ball, P. (2002)
[59] Vogel, S. (1998) in particular chapter 14 on the contrasts between nature and technology.
[60] Kaminuma, T. (1991)
[61] Dupuy, J.-P. (2000)
[62] Boncheva, M., Whitesides, G. (2005) p. 736

mysterious vital force. Today molecular biologists rather look at protein folding or the formation of lipid bilayers as exquisite and optimized mechanisms. Yet self-assembly remains a process of making things through *generation* rather than through *engineering*. Instructions are buit-in the components, instead of being provided by an external program or engineer. To what extent self-assembly could be considered as a technological process of "invention" or "concretization"?

Because of its spontaneity, self-assembly has encouraged the perspective of a new era of technology without human subject. In 1995, Whitesides believed in a future of autonomous machines:

"Our world is populated with machines, non living entities assembled by human beings from components that humankind has made. Our automobiles, computers, telephones, toaster ovens and screwdrivers far outnumber us. Despite this proliferation, no machine can reproduce itself without human agency. In the twentieth century, scientists will introduce a manufactured strategy based on machines and materials that virtually make themselves."[63]

However this autonomy is extremely limited. First, the various techniques of self-assembly developed by chemists anfd biologists over the past decades are not self-replicating techniques. Moreover far from suggesting a process of making things without human intervention the techniques of self-assembly display treasures of ingeniosity: playing with weak forces with energies close to thermmal agitation (such as H-bonding, Van der waals, electrostatic, capillary, hydrophobic and hydrophilic bonding), building templates to grow the aggregate with geomtrical constraints… To be sure nanoscientists and nanoengineers are learning a lot from biology, but they are not simply "mimicking" natural processes. They are using all possible resources from thermodynamics and of chemistry in order to take advantage of molecular interactions for creating order out of disorder, in view of making useful things. So far however, most molecular self-assembly strategies have been confined to static devices, resting on equilibrium at minimum of energy. For inventing "concrete machines" the next step should be making dynamic systems that turn the obstacle of molecular agitation into conditions for the machine to operate.[64] Just as Guimbal designing his turbine chemists and nanoengineers will have to imagine functional structures as "individuals" with their own associated environment.

Conclusion

This paper is only a preliminary attempt at a conceptual clarification of nanotechnology. However it may be useful for the current debates about the so called revolutionary nanotechnology. Drexler claimed that nanomachines would open up a new technological era, but his own "engines of creation" rather suggest the resilience of the old Cartesian paradigm. Although self-assembly and biomimetism may lead to more "concrete machines", most nanomachines currently designed are old wine in new flasks. Dealing with individual molecules does not necessarily entail that a deep revision of conventional engineering methods.

The debates over the control of nanomachines seem to be undermined by a confusion between two distinct notions : Von Neumann's complexity, which would result in undeterministic and uncontrolled machines and Simondon's "technological individuality", which would result in deterministic machines associated with their environment and consequently under better controlled than conventional machines.

In our view, the most immediate dangers do not come from self-replicating nanorobots. They may come from the uncontrollable interactions between the various nanomachines that are being

[63] Whitesides, G. (1995) p. 146
[64] Boncheva, M., Whitesides, G. (2005)

designed and the environment. The relations between machines and their associated environments, between the technosphere and the biosphere have not been seriously investigated and should be paid more attention.

In its ambition to explore the nanoworld by making machines, nanoscience may be seen as the continuation of the chemists' multisecular endeavour for knowing nature through making artefacts. In this respect, tnanoscientists and engineers tend to dissolve the unity of nature constructed by classical mechanism and the grand narratives provided by Newton or Einstein into a multitude of tiny machines. Nanoscientists hold the local but they loose the global view. The famous slogan "shaping the world atom by atom" associated with an image of space is misleading. It diverts the attention from the fact that a jungle of nanomachines is not a cosmos. How those nanomachines fit together and how they operate into a complex system is still unclear.

Acknowledgements
We are very grateful to Dr Christian Joachim, Pr Jean-Pierre Dupuy and an anonymous referre for they critical comments on earlier version of this article. Part of the research for to this essay has been funded by the programme Bionanoethics of the Agence nationale de la recherche (Projet n°NT05-4_44955).

References

Ball, P., 2002: 'Natural Strategies for the molecular engineer', *Nanotechnology*, 13, 15-28.
Ball, P., 2003: 'Nanotechnology in the Firing Line', *Nature*, December, 23, .
Baum, R., 2003: 'Nanotechnology. Drexler and Smalley make the case for and against molecular assemblers', *Chemical & Engineering News*, 81, N°48, 37-42.
Bensaude-Vincent, B., 2001. "The Construction of a Discipline : Materials Science in the U.S.A", *Historical Studies in the Physical and Biological Sciences*,31, part 2, 223-248.
Bensaude-Vincent, B., Arribart, H., Bouligand, Y., Sanchez, C., 2002: "Chemists at the School of Nature", *New Journal of Chemistry*, 26, 1-5.
Boncheva M., Whitesides G.M., 2005: "Making Things by Self-Assembly", *MRS Bulletin*, 30, oct 2005, 736-742.
Breaker, R., 2004: 'Natural and Engineered Nucleic Acids as Tools to Explore Biology', *Nature*, 16 dec 04 p. 838.
Breen, T. L., Tien. J., Oliver, S.R., Hadzic T., Whitesides G., 1999: 'Design and Self-Assembly of Open, Regular, 3D Mesostructures', *Science*, 284 (7 May), 948-951.
Bueno, O., 2004: "Von Neumann, Self-Reproduction and the Constitution of Nanophenomena" in Baird, D. and al. eds, *Discovering the Nanoscale*, IOS Press, 101-118.
Canguilhem, G., 1952: "Machine et organisme" in *La connaissance de la vie*, Paris, Hachette, quoted from the fourth edition Paris, Vrin, 101-128.
Canguilhem, G., 1979: "Le tout et la partie dans la pensée biologique ", *Etudes d'histoire et de philosophie des sciences*, Paris, Vrin, 319-334.
Drexler, K.E., 1981: 'Molecular engineering: An approach to the development of general capabilities for molecular manipulation', *Proceedings of the National Academy of Sciences*, 78, N°9, chemistry section, 5275-78.
Drexler, K.E., 1986: *Engines of Creation*, Anchor Books. Quoted from the 2nd ed. 1990.
Drexler, K.E., 1992: *Nanosystems. Molecular machinery, manufacturing and computation*, Palo Alto, John Wiley & sons.
Drexler, K.E., 2001: 'Machine-Phase Nanotechnology', *Scientific American*, Sept. , 74-75.
Dupuy, J.P., 2000: *The Mechanization of the Mind,* Princeton N.J., Princeton University Press.

Dupuy, J.P., 2004: "Complexity and Uncertainty", in *Foresighting the New Technology Wave*, High-Level Expert Group, European Commission, Brussels.

Fox Keller, E., 1995: *Refiguring Life. Metaphors of 20th century Biology* New York, Columbia University Press.

Guchet, X., 2005: *Les sens de l'évolution technique*, Paris, Editions Léo Scheer.

Joachim, C., Gimzewski G., Aviram A, 2000: "Electronics issuing hybrid-molecular or mono-molecular devices", *Nature*, 408, 541-48.

Joachim, C., 2005: "To be nano or not to be nano ?", *Nature Materials*, 4, February, 105-109.

Jones, R.L., 2004: *Soft Machines*, Oxford University Press, Oxford, New-York.

Kaminuma, Tsuguschika (eds),199: *Biocomputers. The Next Generation for Japan*, New York, London Chapman Hill.

Lafitte, J., 1932: *Réflexions sur la science des machines*, Librairie Bloud & Gay, Paris.

Leigh, D.A., Wong J., Dehez F., Zerbetto, F., 2003: 'Unidirectional rotation in a mechanically interlocked molecular rotor', *Nature*, 424,174-179.

Lehn, J.M.,1985: 'Supramolecular Chemistry: Receptors, Catalysts and Carriers', *Science*, 227, 849-56.

Maurel, M.C., Miquel, P.A.: 2001, *Programme génétique : concept biologique ou métaphore ?*, Editions Kimé, Paris

Merkle, R.,1992: 'Self Replicating Systems and Molecular Manufacturing'. www.zyvex.com/nanotech/selfRepJBIS.html

Neumann, J. Von, "The General and Logical Theory of Automata", *in Cerebral Mechanism in Behavior: The Hixon Symposium*, New York, John Wiley and Sons, 1951

Newman, W., 1989: 'Technology and the Alchemical debate in the Late Middle Ages', *Isis*, 80, 423-445.

Phoenix, C., Drexler, K.E., 2004: 'Safe exponential manufacturing', *Nanotechnology*, 15, 869-872.

The Royal Society and the Royal Academy of Engineering, 2004, *Nanoscience and Nanotechnology : Opportunities and Uncertainties*. London, Document 19/04 ; http///www.nanotec.org.uk.

Saunier, C., 2005: *L'évolution du secteur des semiconducteurs et ses liens avec les micro et les nanotechnologies*, rapport Assemblée nationale (N°566) et Sénat (N°138), Paris, Assemblée nationale, 3 vols.

Sauvage, J.P.: " Les nanomachines moléculaires : de la biologie aux systèmes artificiels et aux dispositifs" . http://culturesciences.chimie.ens.fr/NanomachinesJPSauvage.pdf

Seeman, N.C., Belcher, A.M., 2002: Emulating biology: building nanostructures from the bottom up *Proceedings of the. National Academy of Science USA*, 99, 6451-6455.

Simondon, G., 1989 : *Du mode d'existence des objets techniques*, Aubier, Paris.

Whitesides, G.M., 1995: 'Self-Assembling Materials', *Scientific American*, sept : 146-149.

Whitesides, G.M., 2001: 'The Once and Future Nanomachine', *Scientific American*, Sept : 78-83.

Whitesides, G.M. Wrong A.P. 2006: "The intersection of Biology and Materials Science", *MRS Bulletin*, 31, January 2006, 19-27.

Wood, S., Jones, R.A.L., Geldart, A., 2003: *The social and economic challenges of nanotechnology Economic and Social Research Council* Report available at www.esrc.ac.uk/esrccontent/DownloadDocs/Nanotechnology.pdf

Zhang, S., 2003: 'Fabrication of novel biomaterials through molecular self-assembly', *Nature Biotechnology*, 21, N°10: 1171-78.