SOCIETY FOR PHILOSOPHY AND TECHNOLOGY

Techné:

Research in Philosophy and Technology

Special Issue: Technology and Normativity

Guest Editors:
Ibo van de Poel

Volume 10 Number 1 Fall 2007

Techné: Research in Philosophy and Technology

Editor, Davis Baird
Editorial Assistant, David Stubblefield

CONTENTS

Editorial

Guest Editors, Ibo van de Poel and Peter Kroes

Ibo van de Poel and Peter Kroes, Technology and Normativity 1

Part 1: Technology and Normativity

CARL MITCHAM, In Qualified Praise of the Leon Kass Council On Bioethics 7

LOTTE ASVELD, Informed Consent in the Fields of Medical Technological Practice. 18

JUNICHI MURATA, From *Challenger* to *Columbia*: What lessons can we learn from the accident investigation board for engineering ethics. 35

SVEN OVE HANSSON, Safe Design 54

Part 2: Technological Functions and Normativity

MARCEL SCHEELE, Social Norms in Artefact Use: proper functions and action theory. 65

FRANCOISE LONGY, Function and Probability: The Making of Artefacts. 81

JEROEN DE RIDDER, The (Alleged) Inherent Normativity of Technological Explanations 97

KRIST VAESEN, How Norms in Technology Ought to Be Interpreted 117

Technology and Normativity

Ibo van de Poel
Peter Kroes

This collection of papers, presented at the biennual SPT meeting at Delft (2005), is devoted to technology and normativity. As such, that is a very broad topic, since various kinds of norms play a role in technology. For instance, with science, technology shares the important role of epistemological norms and values; reliability of knowledge claims is clearly of paramount importance in technology. And in various technological practices, such as architecture, aesthetic norms and values play a dominant role. Epistemological and aesthetic norms and values are not, however, the topic of this special issue. It focuses on the role of moral norms and values in technology (the first part, *Ethics and Technology*), and on the normative aspects associated with functions attributed to technical artefacts (the second part, *Technological functions and normativity*). Prima facie, the role of moral norms and values is related to what engineers ought to do, whereas the normative aspects of functions are related to what technical artefacts ought to do.

The contributions of the first part do not address the role of moral norms and values in technology directly. Instead they centre on a meta-issue, namely how to analyze the role of moral norms and values in technological practices. A main theme in all contributions is how ethics of technology should be practised. What approach should be followed? What may we learn from other areas of applied ethics in this respect? The papers of the second part focus on the normative aspects of technical artefacts, in particular on the normative features related to the notion of function. The functions attributed to technical artefacts form the basis for normative claims about these artefacts, for instance about the proper use of those artefacts or about malfunctioning. How are these normative claims to be understood? To what extent are they, for instance, related to moral norms and values or to the norms and values of practical rationality?

Let us briefly point out a development in ethics and engineering that might bring these two clusters of problems closer together. In recent years the issue of the moral agency of technical artefacts is attracting more and more attention.[1] Those who argue in favour of some kind of moral agency consider technical artefacts to be inherently normative: technological artifacts are not taken to be simply inert, passive means to be used for realizing practical ends. In other words, technological artifacts are considered to be somehow 'value-laden' (or 'norm-laden'). These moral values and norms may be explicitly designed into these artifacts, or they may be acquired in (social) user practices. If indeed, technical artifacts are normative in a moral sense, then it may be an interesting opportunity for future research to explore any parallels in our interpretations of the moral and non-moral normative aspects of technical artifacts.

Ethics and Technology

Mitcham's opening paper discusses the approach of the Kass council on bioethics in the USA. Although the council has been heavily criticized, Mitcham argues that there are actually a number of things we can learn from the approach chosen by the council. This approach stands out, Mitcham argues, in three respects. First of all, it involves non-specialists, i.e. people from outside the bioethics community. Secondly, it focuses on bigger ethical issues that new technologies raise and does not only carry out piecemeal or specialized analyses. Thirdly, it refers to human nature as a norm. According to Mitcham, these are also three respects in which the ethics of technology can be improved. Also here, philosophy is not only something for specialists but also for the wider public. The second point means, according to Mitcham, that we should be prepared to talk about technology as a whole and not only about individual technologies. Thirdly, also with respect to technology, we should pay attention to nature as a norm. Mitcham admits that the use of the notion of "nature" in ethical discussions is often unclear or confused. He believes however that the solution is not dismissing the term as such but clarifying what people mean with "nature," especially because the feeling that certain technologies contradict human nature seems to be an important moral concern for many people.

The contribution by Asveld compares the approach of "informed consent" for medical and technological practices. Like Mitcham, she uses approaches and developments in another area of applied ethics, in her case medical ethics, as

[1] For instance, it was one of the main topics of discussion at the workshop on New Directions in Understanding Ethics and Technology, University of Virginia, Charlottesville, October 27-30, 2004.

inspiration for the ethics of technology. In medical practices, informed consent is used for dealing with risks of medical treatment or experiments. The principle serves the goal of protecting the autonomy of the patient: if the patient is fully informed about the risk of treatment, the patient has the free choice to undergo the treatment or not. Asveld argues that technological practices differ from medical practices in three relevant aspects when it comes to informed consent. First of all, the aims are different. Whereas medical practices aim at human health, technological practices aim at human welfare. The goal of health is less controversial and more internal to the practice of medicine than the goal of human welfare is to the technological practice. While in medical practises the desirability of the aim of the entire practice can usually be taken for granted, this is not the case in technology. The second difference is the knowledge of risks. According to Asveld, in medical practices knowledge about risks is less contested, partly because the circumstances of use are more predictable. Whereas in medical practices informed consent can be based on more or less consensual knowledge of risks and therefore focuses on their acceptability, in technological practices discussions about the level of risks and their acceptability cannot be separated easily. Finally, the medical practice is - according to Asveld - more exclusive. With this she means that when people enter the medical practice they already have accepted certain fundamental principles underlying that practice; while in technological practises, which are more ubiquitous, this need not be so.

Murata's contribution criticises the professional approach to engineering ethics. He discusses two interpretations of the disaster with the Challenger. The first one, which can be found in many books on engineering ethics, interprets the disaster as a case in which the risk was known in advance and the accident could happen because engineering judgement was overruled unjustifiably by managers. The second interpretation follows Vaughan's book on the Challenger disaster (Vaughan 1996). In this interpretation, the risks of the Challenger were less clear-cut; moreover, the disaster was not caused by managers overruling engineers but was due to the culture at NASA. This culture had resulted in the "normalization of deviance": risks were not longer perceived as such. Murata believes that the second interpretation is much more plausible. According to Murata, to prevent disasters like that of the Challenger, we should not focus on professional responsibility, but on the inherent unpredictability of technology and the civic virtue of engineers. Engineers should be aware of "normal accidents", i.e., accidents due to the normal procedures in an organization for dealing with technologies and their risks. This requires not just organizational measures but a culture in which engineers are sensitive to the unpredictable. It is here that the notion of “civic virtue” is relevant, i.e. the virtue of caring for others and having regard for their welfare. This virtue is civic and not just professional because it is expected of all citizens.

The contribution by Hansson focuses on safe design. As in Asveld’s and Murata’s contributions, dealing with the hazards and risks of technology is an important theme in his contribution. The focus is, however, different. Whereas Asveld focuses on the acceptability of technological risks and Murata on organizational and cultural measures for minimizing risks, Hansson focuses on design approaches for minimizing risks and hazards. Hansson argues that engineers have an important responsibility for designing safe technologies. This responsibility, however, extends beyond dealing with risks to dealing with uncertainty. Risk refers to the situation in which there is reliable knowledge of the probability of certain undesirable events. In the case of uncertainty, we lack such knowledge. Hansson argues that strategies for safe design are in fact not only strategies for dealing with risk but also for dealing with uncertainty. For example, adding a safety factor to the strength of a construction not only helps in dealing with known fluctuations in loads or material strengths but can also be effective in dealing with unknown failure modes. It is important to be aware of this: replacing current approaches by approaches that only address risks and not uncertainty may lead to more disasters and be ethically unacceptable.

Technological functions and normativity

The part on technological functions and normativity starts with an analysis by Scheele of the role of social norms in artefact use. The use of technical artefacts is, of course, strongly guided by norms of practical rationality, but Scheele argues that more norms are involved, in particular social standards or norms of conduct. Some of these norms are intimately related to the proper functions of artefacts. He argues that proper functions provide “institutional reasons” for use. Proper use of artefacts, viz. use according to the proper function, is embedded in the normative structures of social institutions. These social normative structures are complementary to traditional norms of practical rationality and are a kind of second-order reasons. He claims that proper functions of artefacts provide institutional reasons, which are up to a certain extent similar to what Raz calls ‘exclusionary reasons’. Scheele also observes that institutional reasons may not only give reasons for action, they also provide reasons for evaluating actions. Scheele’s analysis presents a deeper insight into how the interplay of norms of practical rationality and of social norms determines the use of technical artefacts and the evaluation of that use.

The attribution of technical functions to objects has normative implications, but for understanding these implications we need an adequate theory of (technical) functions. In her contribution, Longy addresses a long standing problem with regard to theories of function. It is well known that explaining the (normative) phenomenon of dysfunction (malfunction) has been and still is a real problem for theories of (technical) functions. Because of the phenomenon of dysfunctions it is not possible to simply identify the function to do F with the capacity to do F. But as she rightly observes, we often infer capacities from functions. To solve this problem, she proposes a new theory of functions, of the etiological sort, which is based on a probabilistic relation between having the capacity to do F and having the function to F. This theory, she claims, applies to organisms as well as to artefacts. She argues that the probability of dysfunction may be interpreted in an objective way by distinguishing between considering an object as a physical body and considering it as an artefact. With regard to the object as member of an artefact category, the probability of dysfunction may be taken to be objective because it is causally determined by objective factors. In this way, Longy constructs a probabilistic theory of technical functions that she claims can account for the phenomenon of malfunction.

The normative aspects of technical functions also raise problems with regard to the nature of technological explanations, which is the topic of De Ridder's paper. When designing a technical artifact, engineers are usually able to explain how its function is realized on the basis of its physicochemical properties or capacities. An explanation that purports to explicate this relation between artifact function and structure may be called a technological explanation. There appears to be something peculiar about technological explanations in the sense that a functional property with normative connotations is explained in terms of purely structural (factual) features. De Ridder argues, however, that there is nothing special about technological explanations. He points out that a distinction has to be made between (1) a theory of function ascriptions and (2) an explanation of how a function is realized. The task of the former is to spell out the conditions under which one is justified in ascribing a function to an artifact. A good theory of function ascriptions should account for the normative features of these ascriptions. If that is taken care of by the theory of function ascriptions, then the explanation of technical functions in terms of structures does not pose any special problems. These explanations can pass the buck of normativity to the theory of function ascription. To substantiate his claim, he discusses a particular theory of function ascriptions that in his opinion does account for the normativity of function ascriptions.

In the final paper of this special issue, Vaesen addresses the question what kind of norms are operative in technological practice and how these norms are to be interpreted. He claims that at least two kinds of normativity may be distinguished in technological practice. One kind of norms concerns what engineers ought to do and the other concerns roughly speaking what artifacts ought to do. This claim is controversial in so far as normativity is associated with technical artifacts. According to the standard approach to normativity, namely normative realism, artifacts are denied any kind of normativity, since normativity applies exclusively to human agents. Only human agent normativity is taken to be a genuine form of normativity. Vaesen argues that normative realism is mistaken on this point. Referring to the work of Daniel Dennett and Philip Pettit he shows that it makes sense to talk about artifactual normativity. He claims that his approach can also make sense of human agent normativity. That is an interesting claim, since it implies, *prima facie*, a unified approach to moral and non-moral forms of normativity.

References

Vaughan, D. 1996. *The challenger launch decision.* Chicago: The University of Chicago Press.

In Qualified Praise of the Leon Kass Council On Bioethics

Carl Mitcham

Abstract: This paper argues the distinctiveness of the President's Council on Bioethics, as chaired by Leon Kass. The argument proceeds by seeking to place the Council in proper historical and philosophical perspective and considering the implications of some of its work. Sections one and two provide simplified descriptions of the historical background against which the Council emerged and the character of the Council itself, respectively. Section three then considers three basic issues raised by the work of the Council that are of relevance to philosophy and technology as a whole: the role of professionalism, the relation between piecemeal and holistic analyses of technology, and the appeal to human nature as a norm.

Key Words: Bioethics, Human Nature, Philosophy, Professionalism, Technology

Since its emergence as a well defined field of discourse in the 1970s bioethics has, more than any other form of critical reflection on technology, achieved specific institutional expressions and influenced the practice of technoscience. What follows is an effort to place in historical perspective one of these institutional formations — the President's Council on Bioethics established in 2001 by U.S. President George W. Bush and chaired until late 2005 by Leon Kass — and to consider its implications for philosophy and technology studies. To this end the paper will first review related institutional developments in bioethics, then offer an interpretative description of the Kass Council, before concluding with some general critical comments.

1. U.S. Federal Bioethics Commissions before Kass

From its beginnings bioethics has been manifested not only in academic research and teaching, and in the creation of non-governmental centers, but also in government-related committees or commissions directed toward the formation and implementation of public policy. With regard to academic research and education, the Hastings Center (founded 1969) and the Kennedy Institute (founded 1971) led the way; bioethics journals and bibliographies were established, an *Encyclopedia of Bioethics* was edited (first edition, 1978). With regard to governmental entities, the 1970s and 1980s saw the establishment of federally mandated Institutional Review Boards (IRBs), Institutional Biosafety Committees (IBCs), and Hospital Ethics Committees (HECs) to bring reflective expertise and public consensus to bear on advancing scientific and technological forms of medical research and clinical practice. In the field of biomedicine issues of technology and ethics were given significant theoretical and practical expression.

In many countries there have also existed at the national level bioethics commissions which, in the United States, have been associated with a series of presidential administrations. During the administration of Republican President Gerald Ford (1974–1977) the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, administered by the Department of Health, Education, and Welfare, drafted guidelines useful to both IRB oversight regarding research and HEC guidance of clinical practice. Another recommendation of this National Commission was to establish an Ethics Advisory Board (EAB). During its brief existence from 1978 to 1980, the EAB reviewed issues involving fetuses, pregnant women, human in vitro fertilization (IVF), and initiated a moratorium on human embryo research.

Originally intended to become a standing federal entity, the EAB was disbanded because of perceived overlap with the President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research established in 1978 by Democratic President Jimmy Carter (1977–1981). The Carter Commission, whose report on foregoing life-sustaining treatments led to the development of legal forms for personal directives concerning how one would want to be treated if unconscious and on artificial life support (often termed "living wills"), expired in March 1983 under Republican President Ronald Reagan (1981–1989).

The distinctive Reagan administration contribution was a Biomedical Ethics Advisory Committee (BEAC) to be appointed by a Biomedical Ethics Board (BEB) composed of six Senators and six Representatives. With its creation delayed for more than two years by partisan politics related to the abortion issue, the BEAC officially expired in September 1989 under Republican President George H. W. Bush (1989–1993).

Then in 1995 the administration of Democratic President Bill Clinton (1993–2001) created the National Bioethics Advisory Commission (NBAC). Originally NBAC was tasked with revisiting questions of human subjects research and investigating the proper uses of genetic information. However, after the 1996 cloning of the sheep Dolly, Clinton requested that NBAC direct its attention to the prospects for human cloning. Cloning thus became its first report, which recommended federal legislation to ban somatic cell nuclear transfer to create children. At the same time it argued such legislation should not interfere with less ethically problematic forms of cloning. Before its 2001 expiration, NBAC further produced reports on research involving biological materials, persons with mental disorders that impaired decision-making, and human embryonic stem cells. The policy proposed by the Clinton administration in its final year was to fund embryonic stem cell research but not stem cell line creation.[1]

2. The Kass Council and Its Character

The human embryonic stem cell issue sparked creation of the President's Council on Bioethics by Republican George W. Bush. The issue had actually come up during one of the presidential campaign debates, and Bush had treated it as settled by the National Institutes of Health policy. But after winning the election, Bush revisited the issue, and in August 2001 gave a nationally televised address dealing with stem cell research. Many scientists had been arguing that research using human embryonic stem cells could yield therapies for disabilities associated with Alzheimer's and Parkinson's diseases, diabetes, and spinal cord injuries. But extracting stem cells from human embryos destroys the embryos, which conservative ethicists and many of Bush's Christian supporters found objectionable. The particular question for Bush was whether to endorse the proposed and already ready restrictive Clinton policy or to make the policy even more restrictive. His decision was a slight narrowing that would allow federal funding for research only on those stem cell lines that had already been extracted before the date of his speech. He concluded by announcing that given the importance of this kind of issue he would also create a new presidential council on bioethics to be chaired by Leon Kass in order to further examine the ethical and political issues surrounding all biomedical research.

Kass had earned a BS in biology and an MD from the University of Chicago, followed by a Ph.D. in biochemistry from Harvard University. After some years doing medical research at the National Institutes of Health, he taught in the classics based curriculum St. John's College (Annapolis, Maryland), and then returned to the University of Chicago as a professor in the classics oriented Committee on Social Thought.

Kass's shift from medical research to philosophy was influenced by Leo Strauss, a professor at Chicago and St. John's who sought to revive the philosophical perspective of Socrates, Plato, and Aristotle along with the theological wisdom of the Bible. In promoting these ancient traditions, Strauss was critical of the modern thought that began with Niccolò Machiavelli, Francis Bacon, and René Descartes. He was especially skeptical about the modern project, to be driven by needs rather than guided by ideals, for a new science that would conquer nature. Such a project was likely to eventuate in a willful pursuit of power unconstrained by moral or religious limits that would be ultimately self-destructive. When Kass expressed a version of this skepticism about modern science and technology, he won the attention of those North American political and religious conservatives who themselves harbored suspicions of the modern scientific project as subverting the moral and religious traditions.

In consultation with Kass, Bush appointed 17 persons to the Council. Critics suggested this group was biased by the inclusions of a significant number of political and religious conservatives. Not only did Kass note some hypocrisy in the complaint, since previous federal commissions had excluded representatives of the right to life movement, but it soon became apparent that there was genuine disagreement within the Council. Some members were in fact strong proponents of biotechnology who dissented from Kass's moral criticisms of science. Indeed, in his original executive order creating the Council, Bush had indicated that "the council shall be guided by the need to articulate fully the complex and often competing moral positions on any given issue and may, therefore, choose to proceed by offering a variety of views on a particular issue rather than attempt to reach a single consensus position" (Bush 2001). Kass himself acknowledged that insofar as the Council engaged in serious discussions of the competing human goods animating biomedical research and technology, disagreement would naturally arise because different people often weigh such goods differently. What was important, Kass insisted, was that every serious point of view should be considered as part of a deliberative reflection that might well not reach consensus.

Bioethics scholars also voiced complaints that the Council contained few members who were professional bioethicists. But this exclusion was deliberate, and at the first meeting of the Council, Kass indicated a desire to steer discussion away from the methods and topics that had dominated bioethics as a professional field of academic expertise. "This is a council on bioethics, not a council of bioethicists," he explained. "We come to the domain of bioethics not as experts but as thoughtful human beings who recognize the supreme importance of the issues that arise at the many junctions between biology, biotechnology and life as humanly lived" (President's Council on Bioethics 2002a).

Kass thus guided the Council toward a kind of ethical and political inquiry in which thoughtful persons consider the perennial questions of human life — often using classic texts that illuminate a spectrum of basic alternatives — without expecting to reach closure on the answers. In its initial meeting, Kass actually began by leading a discussion of Nathaniel Hawthorne's short story, "The Birth-Mark," about a scientist who unintentionally kills his beautiful wife while trying to remove a minor blemish. The implication was that the scientific pursuit of power and perfection could, by failing to appreciate the limits of the human condition, turn utopianism into an enemy of the good. Kass thus sought to transform Bioethics Council discussions into something like a seminar on science, technology, and the human condition that would draw not just on technical papers but on the wisdom common to all the humanities, from literature and philosophy to history and religion.

The Council was nevertheless viewed by many as under the thumb of a political agenda. Such a possibility was highlighted when, in early 2004, there was a slight adjustment in Council membership, an event that played out on the pages of the *Washington Post* (see Weiss 2004 and Blackburn 2004, with a response from Kass 2004). More substantively, bioethicist Eric Meslin (2004) argued that the council was not adhering to the spirit of the of Federal Advisory Committee Act of 1972, and bioethicists George Annas and Sherman Elias criticized Kass for leading a "neoconservative bioethics council" that pursued "a narrow, embryo-centric agenda" (Annas and Elias, 2004, p. 19). Although the moral status of human embryos is an important issue, Annas and Elias admitted, such issues as access to healthcare, the commercialization of science and medicine, drug pricing, and bioterrorism also deserve attention. They further charged that neoconservatives such as Kass failed to embrace a global bioethics based on human rights because embryos do not have human status in existing international human rights documents.

In response, Kass suggested that critics had not read the Council's reports carefully enough to see how fair it had been to arguments on all sides of the various debates with which it dealt. Journalists, for instance, had focused attention on some political implications of the Council's recommendations rather than the reasoning supporting them. In many cases recommendations were in fact quite limited, as in reports on using biotechology for enhancing human life, on stem cell research, and on the regulation of reproductive technologies. Although such reports argued Kass's fears about the moral dangers from biotechology, they also regularly acknowledged arguments promoting biotechnology. Indeed, part of Kass's concern has been the attractiveness of the pro-biology arguments. Considering them is part of Kass's way of promoting serious and fair-minded discussion of the deep moral questions raised by modern science and technology.

3. Implications for Philosophy and Technology

During Kass's four-year tenure as chair, the Bush Council on Bioethics experimented with at least three distinctive practices at an important juncture in the historical development of biotechnologies. Since the 1500s the modern technological project has addressed itself primarily to meso- and macro-scales of the external material world; the forces in nature have been progressively harnessed to increase human productivity and to gird the planet with high-speed systems of transportation and communication while subjecting the biosphere to chemical transformation and the risk of nuclear weapons. In the last quarter of the 20th century this project turned its attention toward the micro-level internal workings of the human body and began to imagine the nano-scale reconstruction through design of both life and materiality. At the inauguration of this new phase in the development of modern technology the Kass Council sought to promote thinking (a) that enrolled more than professional bioethicists, (b) that did more than piecemeal or specialized analyses, and (c) that referenced human nature as a norm. Each of these characteristics has implications for understanding the relationship between philosophy and technology.

With regard to stepping outside the professional bioethics community: The professionalization of bioethics may be seen as a version of professionalization of philosophy. In its Greek origins, Socrates criticized philosophical professionalization by distinguishing himself from those who charged money for their teaching and by presenting the practice of philosophy as coextensive with the human good. "The unexamined life is not worth living for humans" (*Apology* 38a). Even in its early modern forms, philosophers such as Bacon and Descartes placed philosophy in the midst of human affairs. The professionalization that took place in the 20th century tended to turn philosophers into experts who work in universities at increasing removes from the public. Something similar took place with bioethicists during the 1980s and 1990s: They became experts who assisted bioscientists and biotechnologists in their work. The Kass attempt to step outside this model of bioethical professionalization thus poses a question for any philosophical attempt to grapple with technology. How much should the philosophical engagement with technology be a matter for experts? To what extent should it instead be an effort to promote critical reflection among a democratic citizenry?

By their very character, questions concerning the proper role for technical expertise in personal and public policy decision making are not answerable by experts alone but require collaboration between non-experts and experts. Moreover, different institutional formations for the development and utilization of expertise have emerged in different countries. For instance, one analysis of the utilization of scientific and technical expertise in Germany noted how such expertise could function in two quite different ways: to improve the quality or effectiveness of political decision making or to rationalize and justify political decisions already made (Brown, Lentsch, and Weingart, 2005). The Kass goal, however, representing what might be called a conservative political agenda for revisiting how science and technology have been used to endorse liberal political interests, has been more to utilize expertise to bring otherwise often marginalized issues and perspectives into the realm of public discussion. The aim has been to revisit or investigate a number of specific policy decisions and the issues at stake as more open aspects of the public agenda.

With regard to the scope of such reflection, whether expert or democratic: Should it proceed on a piecemeal, case-by-case basis, or might it be appropriate to attempt to think technology in general? Coordinate with professionalization has been the rise of disciplinary specialization. Especially in science, analytic distinctions have been drawn between physics, chemistry, biology, and more. Philosophy itself has become more and more fragmented into logic, epistemology, metaphysics, ethics, aesthetics — even 18th or 19th century epistemology or the philosophy of science or pragmatism. In the philosophy of technology there has been an on-going argument about whether there is even such a thing as technology (in the singular), or whether what exists are only technologies (in the plural). On the one hand, division of labor, disciplinarity, and specialization have all been praised for their effectiveness and the ways they have increased the production of knowledge. On the other, there has been increasing recognition that reality itself seldom appears with firm disciplinary divisions and that certain problems are amenable only to multi- or inter- or trans-disciplinary collaboration. Too much focus on rocks alone blinds us to mountains. Thus, in response to the strong tendency in bioethics to focus on case studies of particular issues (the just allocation of dialysis technologies, the ethics of heart transplants, confidentiality in information records keeping, etc.) the Kass Council chose (imitating Martin Luther) to "sin boldly" by conceptualizing biotechnology as a whole and even technology in general. The result, in effect, is to challenge the philosophy of technology to reconsider its contemporary rejections of grand narratives (e.g., Jacques Ellul's *La Technique*) in favor of the manifolds of social constructions and conceptualizations in technologies.

Finally, driven in part by its commitment to public discourse and attempts to talk about technology as a whole, the Kass Council has sought to revive the possibility of referencing nature — especially human nature — as a norm. This is undoubtedly the Council's most problematic stance. It no doubt constitutes in part an extension of Kass's effort in his 1985 book, *Toward a More Natural Science*, to respond to Strauss's admission in the preface to *Natural Right and History* (1953) that the demise of natural law ethics can be traced to a loss of the teleological understanding of nature, which was itself at the core of the modern scientific project that itself turns science into the handmaid of technology. But more importantly, it derives from the fact that in the public realm nature is much more commonly taken as a basis for normative reflection than, say, utility or deontology. Despite more than five hundred years of scientific corrosion acting on the concept of nature, nature to some degree remains a source of awe and a kind of ontological correlate of moral order: "the starry heavens above me and the moral law within me" (Immanuel Kant, *Critique of Practical Reason*, 1788, Akademie vol. 5, p. 161).

It is worth noting, for instance, the persistence of nature as a touchstone in both liberal and conservative traditions of criticism in North America. Political liberals tends to take external nature (the environment) as some kind of good to be preserved, while political conservatives often appeal to social or inner nature (social traditions and human nature) as norms. The former criticize environmental pollution, the latter attempts to alter traditional social orders, including especially religious ideas and beliefs, or proposals to re-engineer human nature. What is significant is that both sides of the political spectrum grant some normativity to some (however attenuated) aspects of nature. Is ethics best served by the prosecution of a philosophical attack on even this residual form of nature, or by some attempt at the sympathetic interpretation if not rehabilitation of such appeals?

Consider from this perspective the Bush Council report *Beyond Therapy: Biotechnology and the Pursuit of Happiness* (2003). In this unexpectedly best selling volume — the popularity of which was paralleled by that of the Council anthology *Being Human* — Kass and colleagues sought to consider broad issues about what it means to be human in the presence of possibilities for the re-engineering not just of the external world but of the inner worlds of human birth, growth, and experience. The 400 page report examines the biotechnological possibilities in both genetic engineering and drug treatment for the parental enhancement of children (chapter 2) and the adult auto-enhancement of, for example, athletic performance (chapter 3). Also considered are the prospects for the transmutation of the experience of aging (chapter 4) and the manipulation of emotion and cognition (chapter 5). In each instance the report expresses special concerns about possibilities for the deformation of humanity not from above by totalitarian governmental use of biotechnology but from below by positive consumer endorsement of the biotechnological satisfaction of desires that a traditional perspective would likely judge to be illegitimate temptations rather than legitimate needs. And precisely because of the admitted inadequacy of the therapy vs. enhancement distinction, *Beyond Therapy* suggests an effort to revive nature as a normative category. Whether and to what degree this is possible remains a fundamental challenge.

These characteristic experiments in bioethical reflection by the Kass Council pose more general challenges in at least two areas. First, there is a challenge to diversify efforts for the critical examination of science and technology — such as those associated with the Ethics and Values Studies program of U.S. National Science Foundation program or the Ethical, Legal, and Social Implications (ELSI) program of the Human Genome Program . One may well ask whether the Kass Council exemplifies approaches that might complement or enhance these other efforts. Second, there is a challenge to the professional philosophy and technology community. Might there be a danger with philosophy and technology studies becoming too professionalized or specialized? Any such questioning will need to include a degree of self-criticism that considers the special responsibilities of a regionalization in philosophy which, more than the philosophy of science or of art, has as part of its heritage public responsibilities and a large measure of ethical concern.

References

Arnhart, Larry. 2005. President's council on bioethics. In *Encyclopedia of Science, Technology, and Ethics*. Detroit: Macmillan Reference. Vol. 3: 1482-1486.

Annas, George J., and Sherman Elias. 2004. Politics, morals, and embryos. *Nature* 431 (7004) (2 September): 19-20.

Blackburn, Elizabeth H. 2004. A "full range" of bioethical views just got narrower. *Washington Post* (March 7): B2.

Briggle, Adam, and Carl Mitcham. 2005. Bioethics committees and commissions. In *Encyclopedia of Science, Technology, and Ethics*. Detroit: Macmillan Reference. Vol. 1: 202-207.

Brown, Mark B., Justus Lentsch, and Peter Weingart. 2005. Representation, expertise, and the German parliament: A comparison of three advisory institutions. In Sabine Maasen and Peter Weingart, eds. *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making*. *Sociology of the Sciences* 24: 81-100.

Bush, George W. 2001. Executive order 13237 (November 28, 2001). Creation of the President's Council on Bioethics. 66 FR 59851. Also available on the President's Council of Bioethics web page: bioethics.gov/about/executive.html.

Caplan, Arthur, and Autumn Fiester. 2005 Bioethics centers. In *Encyclopedia of Science, Technology, and Ethics*. Detroit: Macmillan Reference. Vol. 1: 200-201.

Kass, Leon. 1985. *Toward a More Natural Science: Biology and Human Affairs*. New York: Free Press.

Kass, Leon. 2004. We don't play politics with science. *Washinton Post* (March 3): A27.

Kass, Leon, ed. 2004. *Being Human: Core Readings in the Humanities.* New York: W.W. Norton.

Meslin, Eric M. 2004. The president's council: Fair and balanced? *Hastings Center Report* 34 (2) (March-April): 6-8.

Mitcham, Carl, and Helen Nissenbaum. 1998. Technology and ethics. In *Routledge Encyclopedia of Philosophy*. New York: Routledge. Vol. 9: 280-284.

President's Council on Bioethics. 2002. Transcript of January 17 meeting. Available at http://www.bioethics.gov.

President's Council on Bioethics. 2003. *Beyond Therapy: Biotechnology and the Pursuit of Happiness*. Washington, DC: President's Council on Bioethics.

Strauss, Leo. 1953. *Natural Right and History*. Chicago: University of Chicago Press.

Weiss, Rick. 2004. Bush ejects two from bioethics council. *Washington Post* (February 28): A6.

Informed consent in fields of medical and technological practice: an explorative comparison

Lotte Asveld
Department of Philosophy
Delft University of Technology

Abstract: Technological developments often bring about new risks. Informed consent has been proposed as a means to legitimize the imposition of technological risks. This principle was first introduced in medical practice to assure the autonomy of the patient. The introduction of IC in the field of technological practice raises questions about the comparability of the type of informed consent. To what extent are the possibilities to include laypeople in making decisions regarding risks similar in the technological field to giving informed consent in the medical field and what does this imply for the design and implementation of IC in the technological field? Medical and the technological practice are clearly alike in that both fields are characterized by highly specialized, technical knowledge which can be quite inaccessible to the average layperson. However, a fundamental difference arises with regard to the *aim, knowledge of risks* and *exclusiveness* of the practices in each field. The differences in aim imply that the necessity for each practice is perceived differently by laypeople, thus leading them to assess the respective risks differently. The differences in knowledge of risks arise from the variability in the ways that can be used to describe a given risk. Definition of risk in medical practice is more homogenous in this respect than the risk definition in technological fields. Furthermore, medical practice tends to be more exclusive, leading laypeople immersed in that practice to necessarily embrace most of the fundamental underlying that practice. These differences result in divergent recommendations for the implementation of informed consent in the technological field, basically: there is a need for more extensive procedure and for less decisive authority for the individual.

Introduction

Informed Consent (IC) is a widely used procedure in medical practice where it serves to guarantee respect for the autonomy of the individual. Respect for autonomy requires that an individual can make and enact choices according to her own moral framework. Through the mechanism of informed consent, the patient is given the ultimate authority for deciding the acceptability of a given treatment for herself, after she has been informed by a physician of the risks and benefits attached to the particular treatment. Thus, the patient's autonomy is respected (Faden & Beauchamp, 1986).

The introduction of ICprocedures in the technological field has been proposed as a means to counter some of the ethical deficiencies related to the lack of autonomy for individuals with regard to decisionmaking involving risks (Martin & Schinzinger, 1983). The main aim of introducing IC in the technological field is to give laypersons a greater influence when decisions need to be made about the acceptability of technological risks, rather than, as mostly happens at present, assigning this responsibility entirely to experts.

Although IC has proved applicable in medical practice, its introduction in the field of technological practice would require some amendments. Some salient features of both practices are compared in this paper. The central question is: How do the similarities and differences between medical and technological ICpractice affect the accommodation of individual autonomy in decision procedures involving technological risk?

Technological practice is understood as all those activities that bring forth technological artefacts. Medical practice is understood as activities performed within the boundaries of modern medical science, which center on the human body. Technological practice focuses on the development and production of new artefacts which increase human welfare. Medical practice is concerned with developing artefacts and treatments that are intended to cure human beings and protect their health.

The distinction between the two forms of practice may not be so clear cut as presented here. Medical practice for instance is utterly technological in character, however such overlaps do not invalidate the search for distinctive qualities however. Granted there are similarities between the two fields, the interesting question remains: Where do they differ. More specifically: How do such differences relate to the process of accommodating individual autonomy?

Specific aspects of the two practices that will be used to guide the comparison include: 1. aim, 2. knowledge of risk and 3. exclusiveness. How these aspects are understood is explained below. The aspect of scale is left out of the comparison. Scale is of course one of the most prominent differences between the two practices, as medical practice typically involves only one patient whereas technological practice affects many people at the same time. However, the issue of scale in relation to individual autonomy has been widely discussed. These three aspects may provide interesting insights that have not been considered elsewhere as much as the aspect of scale. Each of the above mentioned aspects will be discussed in in three consecutive parts.

1. Aim

A relevant distinction when comparing technological and medical practice can be discerned in their specific aims. The aim of the medical practice is much more narrowly defined than that of the technological practice. Technology is foremost a means that can be applied to serve a multitude of aims, most of which can be captured under the heading of human welfare. Taking care of people's health is one such aim which may be served by technological practice.

Medical practice, in contrast, serves one clearly identifiable goal, namely to promote human health. This goal is more refined than the sweeping statement human welfare. Moreover, it is a goal that is defined internal to practice. The professional group that practices the art of medicine, also provides the knowledge used to define human health. This is different for technological practice, where engineers aim to serve goals that they do not define by themselves, such goals are defined in communication with clients and regulatory instances that serve to protect the interest of the public at large. Furthermore, engineers rely strongly on other scientific fields when defining the content of such aims as safety, environmental friendliness and economic feasability (Airaksinen, 1994).

This difference implies that the definition and understanding of the aims of medical practice is much more confined to one discipline than that of technological practice. The implications of this for the procedure of Informed Consent will become clear further on.

Additionally, medical practitioners are commonly involved in a practice the main of which is generally unquestioned and the benefits of medical practice are embraced by most people. As Harris & Woods (2001) put it: "We all benefit from living in a society, and, indeed, in a world in which medical research is carried out and which uses the benefits of past research."

Although most people embrace the fruits of medical practice, dissenting voices can still be heard, such as the concerns voiced by Ivan Illich (1976), who questions the alleged achievements of new drugs and research. He points out that many improvements in our health may not be due to better medicine at all, but to better hygiene and food. Moreover, he states, instead of curing people, physicians basically make people (more) ill.

However valid these worries may be, they represent a minority perspective. In Western society in general, there is a strong faith in the beneficence of the medical practices. This strong faith is reflected in what Callahan (2003) describes as the 'research imperative' in the medical context. This imperative refers to the willingness of several actors: industry, government and patient organisations alike, to invest large sums of money in medical research without questioning the effectiveness of such research.

This unreflected faith appears to be much less widely embraced with regard to technological practice. As an illustration: genetic modification as a means to achieve health, i.e. genetic modification of micro-organisms, has remained outside the fierce discussion centering on genetic modification, implying that comparable technologies are assessed differently in different contexts. If it is true that the aim of medicine legitimizes its means more so than for technological practice, this will affect how the procedure of informed consent should be applied in the technological context. People will generally have more and stronger concerns about technology. Since respect for autonomy is the main objective of the procedure of informed Consent, it is necessary to find ways to take these stronger concerns of people with regard to technology into account.

Several reasons exist to suppose the technological practice and accompanying developments are less easily accepted than those of medical practice and the accompanying developments. The first has to do with multiplicity in aims, the second with perceptions of naturalness, the third with perceptions of immediacy and proximity, the fourth with the division of burdens and the fifth with the percieved motives of practitioners.

First, much of resistance against technological development can be explained with reference to disagreements about its aims. There is usually more discussion about the purpose of technological development than about the purpose of medical applications. Technology, in general, can be applied to a wide variety of goals, which might not always seem as pressing as the goal of combating disease. Medicine is a more-or-less one-aim practice as opposed to the multiple-aim practice of technology.

In the resistance to UMTS (3G)-antennas for instance, a technology that offers extended uses for the mobile phone, including watching video's on one's telephone screen, the opponents of UMTS (3G) antennas gave as one of their motivations a lack of need for such a product: "(…) because these UMTS antennas do not serve any other purpose but luxury: the GSMantennas are more than sufficient for the messages-mobiles; the new antennas are nothing but games-antennas for addicted consumers."[1] The intended benefits of this technological development were clearly not recognized as such by these opponents.

Although people might agree that technology spurs progress and that progress is generally thought to be a good thing, the exact implications of what is progress and what is good still leave much to interpretation. Different interpretations may clash. Does progress entail more functions on mobile phones or does it entail less telecommunication? Does progress entail more mobility for more people, or does it entail a healthier environment?[2]

It could be stated that it is not the *aim* of a new technological development that is subject to extensive debate, but rather the *means* available for achieving the goal of the technological 'progress'. So people might agree that alleviation of world hunger is a necessary element of progress, the main disagreement lies in the question whether genetically modified food is an appropriate way to achieve this. However, even if the appropriateness of means is a main cause of disagreement, can the disagreement still be expected to be less intense when the aim of the practice is unambigously defined. 'Health', in this respect, is more straightforward than 'progress'.

[1] Text on pamflet calling for public action against UMTS-antennas, www.stopumts.nl,

[2] Of course, some technologies may be able to combine different interpretations of progress, such as an environmentally friendly car, but often, such aspirations are on a par.

Secondly, as said above, the aim of medical practice is to cure human beings. Let us have a closer look at this aim. What curing actually implies, is a contentious issue, a quogmire into which I will not venture at this point. I will stick to the concept of cure as reflected in Norman Daniels (1985) definition of health: “health is the absence of disease, and diseases (I include deformities and disabilities that result from trauma) are deviations from the natural functional organization of a typical member of a species.” This definition implies that to cure is to restore the natural functional organization of a typical member of a species.

What is important here is the normative connotation cure holds for most people. The aim of medical practice to restore a natural function, as given in the definition above, is easier to accept than the pervasive alterations that are brought about by technological developments, which appear rather to lead to deviations from a natural state instead rather than restoring something to a natural state.

Again, naturalness is a contentious and often culturally biased notion, but nonetheless it is very appealing to those of us in developed societies. It carries with it a reference to a desirable, pure state of being, which is treathened by any kind of modern economical, political or technological progress.

An exception may be the practice of psychiatry where the strive for ‘restoring natural functions’ is less recognizable. It is also in this field that aim and methods used may spark more controversy than other branches of medical practice. While acknowledging this as a problematic instance, I will regard psychiatry as a-typical because the concept of cure and natural functional organization are highly contentious in this medical field. The case of psychiatry does support my thesis that viewing a practice as restoring a natural situation contributes to its acceptability, as a concept of naturalness seems more unattainable in the area of mental illnesses than in other fields of medical practice.

Thirdly, illness brings about a direct pressing need the alleviation of which often becomes a prime objective which dissolves other, more broad-ranging and abstract considerations. The perceived direct need for taking certain risks is stronger in medical practice than in technological practice.

If people are ill, or someone of their loved ones is ill, the prospect of cure may lead them to subject themselves or their loved ones to a system of expert knowledge without questioning the system too much. As Schermer (2001) describes the situation in hospitals: "For patients, there was often not much real choice; a course of action was proposed or prescribed to them that they could either accept or refuse." (p.80). This situation appears to be regarded as unproblematic by most patients. "(…) for many patients, medical decision-making was not something they were very concerned about or wanted to take part in." (p. 85)

People heavily depend on physicians when they are ill and lack the strength or resources to question them.[3]

In the context of medical research, people are often motivated to contribute to a practice, which will eventually benefit themselves or others. In the case of biobanks for instance, people who were interviewed about their motivations to donate blood samples often stated they wanted to help others and the future generation (including their own children). People who did not donate their blood samples, because they thought biobanks might pose a threat to privacy, felt guilty because of this (Haims & Wong-Barr, 2004).

This willingness to contribute to expert health systems, the recognition even of this as a moral obligation, might be explained by the fact that many people have some experience with disease. Most of us have suffered or know someone who has suffered from a disease. The desirability of a healthy life is generally beyond doubt, especially when the negative effects of disease have been witnessed by first-hand experience.

So if we consider the costs and benefits associated with medical applications compared with those of purely technological applications in terms of money spent to achieve a state of well-being for as many people as possible, then medical applications possibly achieve just as much as technological applications. However, the benefits of medical applications may be deemed higher, even if their net result was equal that of technological applications, because they are always bestowed on a specific individual, who is in immediate need of care, whereas the benefits of technological application are more widely spread among a larger group of anonomous individuals, whose needs are seen as less pressing.

Fourthly, aside from this strong appreciation of the benefits of the medical practice, the burdens of technology appear to be much more directly visible. Medicine is often confined to the boundaries of a given hospital and sometimes

[3] This may be different for people who are ill for a long time, and who are not too debilitated. These however are the rarer cases and the main disputes will not take place at moments of Informed Consent.

to the boundaries of a patients' home or an ambulance racing through the streets. Technology in contrast is commonly very visible in society[4], which renders the burdens of such technology much more directly visible.

Paying for your health insurance is much less threatening than having a chemical factory built near to your house. Although in both cases, the individual does not directly benefit from the burdens that are imposed on her, the burdens may be evaluated quite differently as they have quite different characters. The first is a financial burden that is generally easier to carry than a (possibly) life-threatening risk, such as the second. [5]

Put differently, there will generally be less agreement on the necessity for a given technological development than there is for the necessity of a medical development of treatment, as the benefits are easily discernible for medical practice and always pressing, whereas for technological practice, the burdens are more easily discernible.

Fifthly, most healthcare practices and medical researches are usually associated with hospitals and governments, and not with companies, with the exception of pharmaceutical companies. Several people stated in relation to biobanks that they would be less willing to contribute if they were asked by a company for a donation of their DNA material (Busby, 2004: 50). The absence of commercial interests generally contributes to the trust people put in expert health systems. To perceive the interests of the other party as compatible with your own increases the trust one places in the other party (Baier, 1984). This is easier when there is

[4] Not all specific kinds technologies are always visible. Nanotechnology and biotechnology for instance are not easily perceptible for the layperson. However, the point is that people may be better able to witness the negative effects of technology *in general* because they experience technology daily than they are able to judge the negative effects of medicine *in general* as they usually encounter this practice only in very specific instances. Circumstances, moreover, in which a direct need for medicine is felt.

[5] Additionally it can be stated that the health insurance burden is at least shared throughout society, whereas the burden of the chemical installation is directed at one specific geographical area. It is much more difficult to accept this burden, when the benefits are not directly visible, than it is with the burdens of medical practice. The general observation is that the benefits of medical practice are usually aimed at specific individuals while the burdens are evenly distributed in any society. In contrast, in technological practice, the benefits are often accessible to society at large, whereas the burdens are imposed on a limited group of people.

lack of commercial interest. The development of technological artifacts often involves commercial interests.

In conclusion it can be stated that medical risks are usually thought to be more acceptable since they are typically considered to be legitimated by the aim of medical practice for the various reasons mentioned above.[6] The IC-procedure in the medical context therefore serves mainly to protect the patient from deception and coercion (O'Neill, 2002: 97) and does not require discussing the aim of the proposed treatment or experiment.

People will in general be more concerned about the choices that predate the development of technological developments. The fact that such choices are more often of importance to them, justifies their inclusion in the procedure used in the field of technology to gain informed consent. To exclude such issues from the procedure of informed consent will undermine their autonomy, as individuals will be denied the opportunity to make and enact choices according to their own moral framework. If it can be expected that such choices will elicit little discussion and little concern, the conclusion can be drawn that the choices as made by the experts alone, will overall coincide with the choices laypeople would deem most desirable. As it is however, laypeople appear to have strong concerns about the (alleged) necessity of technological development.

This is far less problematic, though not completely unproblematic, in the case of medical developments as laypeople have fewer reasons to question necessity in this area since the aim of medical practice seems to justify for most people most of the risks associated with this practice. However, even if the aim of a technological development was defined straightforwardly and widely embraced, a discussion about the acceptability of the risks involved is still more likely to occur for the same situation in the medical practice.

2. Uncertainty

[6] However, it would be false to state that medicine has not experienced some of the distrust towards its institutions that has characterized the scientific and technological practice. The aim of medicine (cure) does not always legitimize its means or its methods to everyone, as is shown by the growing number of people who turn to alternative health practitioners. Such debates are however usually not conducted in the process of informed consent, though possibly they should be.

Deciding on the acceptability of a certain risk involves two different fundamental issues. The first is A) how a risk is identified and estimated. The second is: B) how a risk is evaluated (Shrader-Frechette, 1991). For the technological practice, both issues have characteristics that have led to the increasing inclusion of lay perspectives in the risk decisions process. For medical practice, only the latter issue is considered in the demand for input from the layperson or patient (Faden & Beauchamp, 1986). Not only is her input required, she is considered to be the sole authority on this matter.

With regard to B, evaluation of a risk, this is basically a *moral* issue. When scientifically trained experts or dedicated policymakers or physicians are given the sole authority to decide on such questions, their specific moral frameworks alone should determine the answer to such questions. Such a decision structure excludes other moral considerations, such as those that laypeople may hold, thereby undermining their autonomy. In matters of purely technical or scientific character, experts or dedicated policymakers may legitimately provide the relevant answers without undermining the autonomy of laypeople, as these questions typically have only one, or a limited number of adequate answers, which the experts will most likely be able to find. This is in contrast with the issue of evaluating the risk: judging whether it is worth taking that risk, here the layperson offers valuable expertise as this is not a technical but a moral issue.

This brings us to issue A: many decisions about risk suffer from lack of certainty or incomplete information. Scientific knowledge often fails to provide conclusive evidence to establish the nature of a given risk. Risk-assessors therefore necessarily rely on assumptions of a non-epistemic, moral, social, economic, kind to determine what constitutes a risk (Shrader-Frechette 1990, Fischoff 1981, Wynne, 1980). Such assumptions are primarily based on a specific worldview which is not scientifically falsifiable. The presence of such assumptions in risk-assessment is inevitable. They should not be regarded as problematic for the field of risk-assessment as knowledge production isn't hindered but they do cause substantial uncertainty. Variance in such assumptions leads to variance in the outcomes of the estimation of a risk.

These assumptions can not be eliminated or reduced. There is reason to suppose however, that they cause less uncertainty about risks in medical practice than in technological practice.

There are two main reasons to suppose this is the case: one, in medical practice new products are extensively tested in controlled environments and two, and related, the application of a new product is very narrowly defined. Qualified professionals may only apply some products; others, for example, can only be taken by people who have obtained prescriptions, based on need, from qualified professionals.

In contrast, the release of most technological products in society is not characterized by the qualities of medical research and application: there are no controlled circumstances. Of course guidelines exist to guarantee safety and products will be tested before they are released onto the market which offers some means of control over the effects the technology will have on society. However, as Van Gorp (2005) describes, for new, radical designs especially such regulatory frameworks are often inadequate.

The main instruments of control for technological products are actual risk-assessments, which suffer from uncertainties. Many of the uncertainties in technological risk assessment arise out of differences in assumptions present in the models applied. These assumptions may involve the way a technological artifact is used, under what circumstances, what kind of events might cause it to malfunction, how it will affect its environment. Such assumptions are very likely to diverge considerably among risk-assessors, since they cover a whole range of environmental, human and technological qualities and reactions that are hard to predict, either because of lack of knowledge or due to sheer complexity.

In contrast, the medical context appears to be much more predictable. This is not to say that in the medical context, no surprises ever occur or that controversies never arise. The likelihood is just much smaller for two reasons. Firstly, knowledge of risks has mainly remained within the technical-medical discourse, confined to medical laboratories and institutions whereas with technology and its attached risks are much more out in the open. The assumptions that underlie the descriptions of risks vary in medicine to a much lesser extent than the risk assumptions made in the technological practice. Second, the context of medical practice is much easier to control in general than the wide-ranging (indefinite) context of technological practice; this widens the variation in assumptions.

This is not to say that medical practice is a necessarily a lot safer than technological practice. The kind of risks in medical practice however, can be said to be easier to describe. This implies that when a patient or a research subject decides on the acceptability of a risk attached to a certain treatment or experiment, the knowledge of the risks involved has been presented in a more homogenous manner as is the case for technological risks. This is true for treatment as much as it is for experiments.

The point is about the nature of our knowledge of risk. The social institutions and relations that underlie each of the practices are fundamental to understanding the generation of knowledge of risk. The social practice of medicine is much more narrowly defined than that of technological practice. This aspect does not necessarily relate to what the risks amount to precisely, it concerns primarily the way the risks are interpreted and described.

Therefore, the patient or the research subject evaluates a specific risk and its estimation will not give much room for multiple interpretations. The possible description of risk is a lot narrower and therefore less debatable in the medical practice than in the technological practice for the reasons mentioned above.

Presumably then, decisions made in medical practice are understood as relating solely to the last stage of risk management: evaluating the risks. The issue of establishing the risk is not so much an issue, therefore the focus is on evaluating the risk. The autonomy of the individual is deemed to be a legitimate concern in this stage of evaluating the acceptability of a certain treatment in medical practice, but not in any other stage (estimation or identification of a risk) as this is a stage that is accessible to experts only.

This gives us reasonable assurance that the judgment of the layperson will be aimed primarily at the *moral* issue of evaluating the risk. In this area, the layperson is usually considered the ultimate authority as her health is at stake and she is the one who knows best what risks she is willing to take to consolidate her health.

In contrast, in technological practice, the realization that the establishment of a risk should, to some extent, be opened up for the input of laypeople, has arrived at the forefront of the consciousness of experts and policymakers. This has led to the inclusion of laypeople in this specific stage of risk-assessment. In many European countries, laypeople are being increasingly invited to participate in decisions regarding the acceptability of technology, to take part in Participatory Technology Assessments (PTA). These assessments resemble the procedure of gaining informed consent in medical practice in that the participants are first informed about the technological development at stake and its accompanying risks, and then get a change to form an opinion about this technology (Europta, 2000, Asselt, v. et al., 2001).

However, even if the fallibility of the expert is recognized, a scientifically trained person might still do a lot better at estimating a risk than a layperson. So even though the input of laypeople is valued very strongly, their perspective on matters is not binding, it is taken as a valuable addition and nothing more. Their input primarily helps experts to overcome the confines of their own limit visions. That this input is solely of an advisory character is a salient difference with the status of the input of laypeople in medical practice, where it is binding.

As the evaluation and estimation of a risk are intertwined more strongly in technological practice than in medical practice where the knowledge of risk is more diffuse, it is more accepted that the layperson has a binding say in the matter of evaluation of the risk in the medical practice. This stage is severed from the other stage of estimating the risk in the medical risk, so the layperson can be considered a true and sole authority; this is not the case for the technological practice.

3. Exclusiveness

Another important aspect when comparing medical and technological practice on their respective suitedness to accommodate informed consent is the moment in development when laypeople can voice their concern. A salient aspect of medical practice in relation to informed consent is that once people are asked to give their consent, they have already crossed a certain threshold. They have already accepted the premises on which medicine is founded. Otherwise they wouldn’t go to a particular physician; otherwise they would not participate in a particular research project. Informed consent in medical practice is mainly a safeguard against abuse; it does not offer a forum to discuss more fundamental issues such as the appropriateness of the method used or deployment of resources. This stage is passed over at the moment Informed Consent is given in the medical practice.

Additionally, a strong boundary exists between what are considered to be legitimate means and practices and what are not. Alternative medicine is a clearly distinguishable medical system. Patients who turn to these practitioners can do this at their own risks and they may be considered less rational for taking such a course.

Medicine is a closed system to a larger degree than technology: the conventional medical institutions are very recognizable and one is either an insider or an outsider. This becomes clear for instance in the fact that not everyone can take part in the medical activities whereas technological development can be undertaken by anyone who is willing to get involved. Medicine is a profession with an internal judicial system, which implies that doctors can be expelled from their professional group if they are convicted of misconduct. In the United States, such a system also exists for engineers but only for engineers not working in industry. The European Union does not have such a system. Engineers do not require a special license in the EU to demonstrate their trustworthiness to third parties.

The closedness of the medical system is further strengthened by the professional loyalty that exists among physicians. Loyalty to one's colleagues and teachers also forms the first part of the Hippocratic Oath. This loyalty makes public discussions of controversies and the reporting of poor practice less likely than in the technological practice, where such loyalties may exist, but only implicitly.

In technological practice the different means used to achieve similar goals cannot so easily be judged solely by the identity of the institutions that propagate them. A wide variety of actors produces technological artifacts, for example companies, universities, inventors and research institutes. There is no formal system to distinguish between the actors if they are not universities. There is however a similar effort as that found in the medical practice to distinguish reliable and unreliable *knowledge* using the distinctive qualities of institutions that assess and publish such knowledge. However, the boundaries between conventional, scientifically sound knowledge and practices are less clear-cut in technological practice. The opponents of UMTS technology for instance, put forward numerous scientific publications, which indicate electromagnetic radiation emanating from UMT Antennas harm human health. Public advisory bodies state however that such evidence is flawed and unreliable. This might be the case, although for the outsider both publications that do not find any negative effects on health and those that do, seem very similar. [7]

[7] This may also be the case for vaccines, where laypeople mobilize medical knowledge to show that vaccines might be harmful to children. Vaccines are not a typical medical case, since the people who are vaccinated are not ill. Thus the benefits of the treatment are less clearly perceptible. This might explain why this is a medical area where fierce discussions arise.

The above described difference can be (partly) explained by the points raised in the section about aims. Medical professionals exercise and define the aim of their practice: namely promoting human health. They do not need to rely on external sources of knowledge. This is different for technological practice, where engineers have to rely on external sources to explain the exact nature of a very broad, almost non-exclusionary aim: to promote human welfare. The 'self-sufficiency' of medical practice explains why it is more of a closed social system of knowledge production than technological practice.

Inclusion in medical institutions requires some concurrence with the basic premises these institutions are founded on. This is true for professionals as for patients. The practice of informed consent in medical practice will therefore never be directed at the basic assumptions underlying this specific practice, as these are taken to be commonly shared by everyone entering into this practice. However, in technological practice this will not be the case as the foundations of this practice are much less exclusively defined. On the contrary, the foundations are necessarily vaguely defined, and may be constantly open to revision. There is little legitimate basis to claim that the definition of technological foundations is limited to professionals alone.

4. Conclusion

Three main conclusions can be derived from the above. One, the concerns of individuals does not require the same elaborate attention in medical practice as in technological practice. This is basically because the aim of medical practice is less multifarious and less open to interpretation than that of technological practice.

Two, descriptions of the risks in medical practice are more narrowly described and understood in specific (medical) jargon. Moreover, they arise in a more controlled setting where the risks are easier to foresee. This assures that the stage of risk-estimation and risk-evaluation are separated which legitimizes the position of the layperson as sole authority for risk assessment in medical practice.

Lastly, the social institute of medical practice maintains a strong external-internal division. Inclusion in medical institutions requires corroboration with the basic premises such institutions are founded on, this is true both for professionals and for patients. This is another feature of medical practice that makes disputes less likely and less frequent than in technological practice. The introduction of informed consent procedures in technological practice requires a different set-up and should have a different scope than in medical practice, this is necessary to accommodate the differences discussed above.

There is a need for more extensive debate and opportunities to discuss fundamental issues in technological practice than there is in medical practice. This is to ensure that disputes about the proposed aim of technological developments and any divergent perceptions of risks are acknowledged and articulated. This is a first step in respecting the autonomy of the individuals involved.

It is less problematic to let the judgment of the individual be binding in medical practice. This is because medical practice does not allow much room for divergence in interpretations of risk. The issue at stake is therefore solely the moral evaluation of the individual, not a perception of the risks at stake. It is clear that the individual is a legitimate authority on the first issue, but with regard to the last issue, this is more problematic. In technological practice, where both stages are less easily to separated, the input of laypeople is welcomed, but not considered to be decisive.

This conclusion was reached mainly by taking the characteristics of both practices as given and constitutive for the needs and autonomy of individuals immersed in medical and technological practices. It may alternatively be suggested that the characteristics of these practices need to change if the autonomy of individuals is to be truly respected, but that suggestion, however interesting, is outside the scope of this paper.

Bibliography

Brody, B. A. 2001. A historical introduction to the requirement of obtaining informed consent from research participants. In *Informed Consent in medical research*, edited by L. T. Doyal, J S. London: BMJ Books.

Busby, H. 2004. Blood donation for genetic research: what can we learn from donors' narratives? In *Genetic databases: socio-ethical issues in the collection and use of DNA*, edited by R. C. Tutton, O. London: Routledge.

Daniels, Norman. 1985. *Just Health Care*, Cambridge: Cambridge University Press.

Faden, R, & Beauchamp, Tl. 1986. *A history and theory of Informed Consent*. New York: Oxford University Press.

Fischhoff, B. et al. 1981. *Acceptable Risk*. New York: Cambridge University Press.

Haimes, E. & Whong-Barr, M. 2004. Levels and styles of participation in genetic databases: a case study of the North Cumbria Community Genetics Project. In *Genetic Databases: socio-ethical issues in the collection and use of DNA*, edited by R. C. Tutton, O. London: Routledge.

Harris, J & Wood, S. 2001. Rights and responsibilities of individuals participating in medical research. In *Informed Consent in medical research*, edited by L. T. Doyal, J S. London: BMJ books.

Kaye, J. 2004. Abandoning Informed Consent: the case of genetic research in population collections. In *Socio-ethical issues in the collection and use of DNA*, edited by R. C. Tutton, O. London: Routledge.

Kluver, et al. 2000. Europta: European Participatory Technology Assessment. Copenhagen: The Danish Board of Technology.

Martin, M W & Schinzinger, R. 1983. *Ethics in Engineering*. New York: Mac Graw Hill.

O'Neill. 2002. *Autonomy and Trust in Bioethics*. Cambridge: Cambridge University Press.

Schermer, Maartje. 2001. *The different faces of autonomy, a study on patient autonomy in ethical theory and hospital practice*. Ridderkerk: Ridderprint B.V.

Shrader-Frechette, K. 1991. *Risk and Rationality: the philosophical foundations for populist reforms*. Berkeley: University of California Press.

Wynne, B. 1980. Technology, risk and participation: on the social treatment of uncertainty. In *Society, technology and risk assessment*, edited by J. Conrad. London: Academic Press.

From *Challenger* to *Columbia*: What lessons can we learn from the report of the Columbia accident investigation board for engineering ethics?

Junichi Murata
University of Tokyo, Department of History and Philosophy of Science,
3-8-1 Komaba, Meguro-ku, Tokyo, Japan
phone: +81 3 5454 6693, fax: +81 3 5454 6978
email: cmurata@mail.ecc.u-tokyo.ac.jp

Abstract: One of the most important tasks of engineering ethics is to give engineers the tools required to act ethically to prevent possible disastrous accidents which could result from engineers' decisions and actions. The space shuttle Challenger disaster is referred to as a typical case in almost every textbook. This case is seen as one from which engineers can learn important lessons, as it shows impressively how engineers should act as professionals, to prevent accidents. The Columbia disaster came seventeen years later in 2003. According to the report of the Columbia accident investigation board, the main cause of the accident was not individual actions which violated certain safety rules but rather was to be found in the history and culture of NASA. A culture is seen as one which desensitized managers and engineers to potential hazards as they dealt with problems of uncertainty. This view of the disaster is based on Dian Vaughan's analysis of the Challenger disaster, where inherent organizational factors and culture within NASA had been highlighted as contributing to the disaster. Based on the insightful analysis of the Columbia report and the work of Diane Vaughan, we search for an alternative view of engineering ethics. We focus on the inherent uncertainty of engineers' work with respect to hazard precaution. We discuss claims that the concept of professional responsibility, which plays a central role in orthodox engineering ethics, is too narrow and that we need a broader and more fundamental concept of responsibility. Responsibility which should be attributed to every person related to an organization and therefore given the range of responsible persons, governments, managers, engineers, etc. might be called "civic virtue". Only on the basis of this broad concept of responsibility of civic virtue, we can find a possible way to prevent disasters and reduce the hazards that seem to be inseparable part of the use of complex technological systems.

Key Words: engineering ethics, risk, safety, space shuttle accidents, civic virtue

One of the most important characteristics of technology is that we can use it to produce certain instruments that can then be used to lighten our work loads, and/or produce safer working conditions for ourselves. It is also well known that the meaning of technology cannot be reduced to the role of instrumentality (Tenner 1996). For example, during the process of production, and while using technology, unintended situations sometimes arise which can be considered a source of creativity but can also lead to technology failures and accidents. How can we interpret this unpredictable and unmanageable aspect of technology? I think this problem is pivotal to the philosophy of technology.

In the view of technological determinism, the processes of technological development and the introduction of a technology to a society are seen in hindsight. This hindsight allows us to interpret them as the processes dominated by technological rationality and efficiency and the unpredictable and unmanageable aspect of technology remains out of focus. In contrast to this deterministic view, the social constructivist approach focuses on technological unpredictability and unmanageability and finds that these aspects provide interpretative flexibility and a chance for users of a technology to take the initiative to develop the technology in a new direction. Thus, our perspective on the philosophy of technology depends on how we characterize these aspects of technology, or on which facet of these aspects we focus (Murata 2003a; 2003b).

A similar situation can be found in discussions of the ethics of technology. One of the most important tasks of those dealing with the ethics of building /using technology is to clearly define methods that can be used to predict and control the process of technology development and by so doing minimize the potential for the new, redeveloped technology to cause harm. However, if an unpredictable and uncontrollable character is essential for the processes of development and use of technology, those dealing with the ethics of technology will be confronted with an apparently contradictory task, i.e. that of predicting and controlling an unpredictable and uncontrollable process (Murata 2003c).

In spite of these circumstances, in discussions of "engineering ethics," a topic which has recently become very popular in the field of applied ethics, this issue has not been sufficiently emphasized as a fundamental and central problem of the field, although it is sometimes touched on. It is common in orthodox textbooks of engineering ethics to describe examples of difficulties engineers meet in their workplaces in such a way that what engineers as professionals must do is clear from the beginning. The difficult ethical problem comes later when the question of how an engineer must realize a task comes up against various disturbing factors arising from circumstances outside the particular technological domain in question. If we regard the problems raised in the field of engineering ethics in this way, the essential character of uncertainty is neglected and the contradictory character of the task of engineering ethics is left out of consideration.

In this paper, I would like to consider the status and the significance of this problem of uncertainty in the field of engineering ethics, taking two famous examples of technological disaster; the two space shuttle accidents. How can ethicists deal with the unpredictable, uncontrollable and creative character of technology? And: what is important in this field, what do engineering ethicists need to do to deal with this problem? These are the questions that I will address in the paper.

I will start with an analysis of the report of the Columbia accident investigation board (Report 2003). This report clearly demonstrates that the essential cause of the accident was to be found not in the failure of individual decisions of engineers or managers, as is usual in the orthodox view of accidents, but rather in structural features, such as the history and culture of NASA, in which the methods used to deal with various uncertainties, dangers and risks are institutionalized and at the same time the organizational sensitivity to possible dangers has gradually been paralyzed. If we take this interpretation of accidents seriously, engineering ethics, which is based on such concept as “professional responsibility” in the narrow sense, is insufficient, as this narrow definition focuses too much on individual decisions and professional actions and overlooks the role of history and culture as an implicit background to each action within an organization. In contrast to the usual perspective taken on engineering ethics, in this paper I focus on a much more fundamental and wider dimension of ethics, in which ethical virtue such as sensitivity to possible danger plays a central role as a cultural element. This virtue can be called “civic virtue”, as it can be attributed to “professionals” and also to every person involved in a technological system. Only when we choose to look at responsibility in a broad sense, we can find a possible way to cope with the apparently contradictory problem of preventing a hazard, something that is inevitable in complex technological systems, such as the space shuttle program. This is, I think, one of the most important lessons we must learn from the report of the Columbia accident investigation board.

1. *Columbia*: report of the accident investigation board

On February 1, 2003, the space shuttle Columbia disintegrated in flames over Texas a few minutes before its scheduled landing in Florida after its 16 day mission in space. This flight, called STS-107, was the space shuttle program's 113th flight and Columbia's 28th. This was the second disastrous accident in the history of the flight of the space shuttle, which began in 1981 with the first flight of the same orbiter Columbia. The first accident was the explosion of the Challenger shuttle, which exploded just 73 seconds into its launch on January 28, 1986.

Immediately after the Columbia accident, the "Columbia accident investigation board" was organized to conduct a widespread investigation. In August, 2003, about seven months after the accident, the voluminous report of the board was published. In the report we find analyses of many kinds of documents and debris and a far-reaching examination of the organizational problems of NASA, which are rooted in its long history.

The report clearly identifies the physical cause of the accident. The physical cause was found to be a breach in the thermal protections system on the leading edge of the left wing of the orbiter, caused by a piece of insulating foam which separated from the external tank 81.7 seconds after launch and then struck the wing. During re-entry this breach allowed superheated air to penetrate through the leading edge insulation and then progressively melt the aluminum structure of the left wing, resulting in a weakening of the structure until increasing aerodynamic forces caused loss of control, failure of the wing, and break up of the orbiter (Report 2003: 9,49ff.).

The board of commission set up to seek the cause of the Columbia disaster did not rest with looking for the physical causes of the disaster, it also looks widely into other areas, especially in organizational factors. This report takes a very critical stance towards NASA's fundamental attitude to the shuttle program in general, an attitude arising from the history of NASA and which is rooted in its culture.

> "The Board recognized early on that the accident was probably not an anomalous, random event, but rather likely rooted to some degree in NASA's history and the human space flight program's culture." (Report 2003: 9, 97, 177)

The report makes it clear that the fact that a lot of foam debris had struck the Orbiter in an unusual way right after the launch was not overlooked by the people in NASA but rather was focused on by many engineers from the very beginning phase of Columbia's flight.

As soon as Columbia reached orbit, engineers belonging to the photo working group began reviewing liftoff imagery recorded by video and film cameras and noticed that a large piece of debris from the left bipod area of the external tank had struck the orbiter's left wing: because they did not have sufficiently resolved pictures to determine potential damage and had never before seen such a large piece of debris strike an orbiter so late in ascent, the engineers decided to ask for ground-based imagery of Columbia, requesting shuttle program managers to get in contact with the defense force (Report 2003: 140f.).

Having heard that a large piece of debris had struck the orbiter's wing and having been anxious about the possibility of a disaster resulting from this fact, engineers belonging to various sections began to analyze and discuss the issue. They even constituted a debris assessment team and continued to work through the holidays. They also tried to obtain imagery of the current situation of the left wing of the orbiter in flight, informing managers about their concern and their requests to obtain a good image of the wing damage. In all the engineers attempted three times to get its imagery, however, each of these attempts was unsuccessful, because they were not made through the formal hierarchical route and the request was ultimately declined by a chief manager of the mission management team.

Some engineers were frustrated by this result, but they could not make the mission management managers listen to their concern. In the formal meeting led by the mission managers, the engineers could not demonstrate the hazard presented by the impact of the debris and were unable to persuade managers to take action because they did not, and could not, acquire the right kind of detailed information. Mission managers, whose main interest lay in keeping the flight on schedule, did not pay much attention to this foam strike. Above all they relied on their presupposition that many previous flights had been successful in spite of debris strikes and that in this sense the debris strike could be considered an "in family" and "turnaround" issue and in this sense an "accepted risk" and not some event which had significance for "safety of the flight". In this situation engineers found themselves in "an unusual position of having to prove that the situation was *unsafe*—a reversal of the usual requirement to prove that a situation is *safe*" (Report 2003: 169).

The Report focuses attention on various organizational and cultural factors as the main causes of the accident, i.e. reliance on past success as a substitute for sound engineering practices, organizational barriers that prevented effective communication of critical safety information and stifled professional differences of opinion, and so on (Report 2003: 9, 177).

> "In the Board's view, NASA's organizational culture and structure had as much to do with this accident as the External Tank foam. Organizational culture refers to the values, norms, beliefs, and practices that govern how an institution functions". (Report 2003: 177)

Reading through the report, we cannot help but notice the similarities between the story of the Columbia accident and that of the Challenger. In fact, in many places the report made a comparison between these two accidents and found in the later accident "echoes of Challenger".

> "As the investigation progressed, Board member Dr. Sally Ride, who also served on the Rogers Commission, observed that there were "echoes" of Challenger in Columbia. Ironically, the Rogers Commission investigation into Challenger started with two remarkably similar central questions: Why did NASA continue to fly with known O-ring erosion problems in the years before the Challenger launch, and why, on the eve of the Challenger launch, did NASA managers decide that launching the mission in such cold temperatures was an acceptable risk, despite the concerns of their engineers?" (Report 2003:195)

Reading the report, we can ask exactly the same question concerning Columbia: Why did NASA continue to fly with a known debris strikes problem in the years before the Columbia launch? And: Why, during the flight of Columbia after the debris strike, did NASA managers decide that the reentry of Columbia with the strike damage was an acceptable risk, overrunning the concerns of their engineers?

What lessons did the managers and engineers in NASA learn from the Challenger accident? Or had they in fact learnt nothing from the Challenger accident?

Our questions or doubt only becomes intensified when we consider the status of the Challenger accident in the field of engineering ethics. In almost every textbook of engineering ethics we find the story of the Challenger accident as a typical case in which the ethical problems of engineering can be seen and from which something can be learnt. Every student, post Challenger, who has attended a course in engineering ethics remembers at least the word "O-ring", the sentence one of the managers of Morton Thiokol said to his engineer colleague at the decisive moment of the decision making, "take off your engineering hat and put on your management hat", and the engineers' hero, Roger Boisjoly, who stuck to his professional conscience until the last moment.

The time between Challenger and Columbia saw the inception and rise of a new discipline, that of engineering ethics. During this time everyone working in the field of technology began to hear about various ideas being discussed in the field of engineering ethics. Practitioners and students should have become more conscious than before that safety was a high priority objective in their professional field. For example, many professional groups expected their members to work according to various codes of ethics, first either setting in place a code or overhauling a code if a certain field already had a code of ethics. Looking at these circumstances, we are lead to ask what kind of role can engineering ethics have in a real work place. In the face of the fact that almost the same accident can happen, precisely in the place where the lessons of the first accident should have been learnt, can we still argue that engineering ethics has a meaningful role? What is lacking? And: What is wrong in orthodox engineering ethics?

In order to tackle these questions, I would like to examine how the accident of Challenger is dealt with in popular textbooks.

2. *Challenger*: two stories

We have now at least two different versions of Challenger accident. One is orthodox, on which discussions in orthodox textbooks of engineering ethics are based. The other is a revisionist version, which seems to be more realistic but difficult to use for engineering ethics. In comparing these two stories, I hope to find some hints on how to revise orthodox engineering ethics.

(1) Story 1, a paradigm case of engineering ethics

It is widely recognized that engineering ethics should be classified as professional ethics, in other words, that because engineers have a special knowledge and influential power as engineers they have a special responsibility to prevent dangerous results caused by their actions. In this sense, engineers must be much more ethically careful when they act as engineers than when they act in everyday situations. Various concepts and issues belonging to engineering ethics are characterized under this presupposition. For example, various professional codes of ethics of engineers are interpreted as rules which explicitly determine what engineers have to do to fulfill their special professional responsibility as engineers. Various concepts, such as honesty, loyalty, informed consent or whistle blowing, are considered to have a similar role as that of codes of ethics, i.e. they should be used to give people guidance on how to decide and act, in order to fulfill their professional responsibility as engineers (Harris et al. 2000: chap.1 and chap. 6; Davis 1998: chap. 4; Johnson 2001: chap.3).

At first sight the Challenger accident seems to give us an impressive example that we can use to learn what it is like to act ethically as a professional engineer in a concrete situation.

The fundamental presuppositions of the story in the orthodox version are given below:

> (a) Engineers knew the problem concerning O-ring very well."Chief O-ring engineer Roger Boisjoly knew the problems with the O-ring all too well. More than a year earlier he had warned his colleagues of potentially serious problems." (Harris et al. 2000: 4f.)
>
> (b) Although the data given on the eve of the launch were incomplete, it was clear that a correlation exists between temperature and resiliency of the O-ring. "The technical evidence was incomplete but ominous; there appeared to be a correlation between temperature and resiliency". (Harris et al. 2000: 4f.)
>
> (c) With respect to the value evaluation there was a clear difference or conflict between engineers and managers. Engineers regarded safety as more important than schedule or profit, and managers prioritize these issues in a reverse order. "Turning to Robert Lund, the supervising engineer, Mason directed him to "take off your engineering hat and put on your management hat." (Harris et al. 2000: 4f.)
>
> (d) Boisjoly is considered to be a role model for engineers, although the result of his action was unsuccessful. "It was his *professional* engineering judgment that the O-rings were not trustworthy. He also had a *professional* obligation to protect the health and safety of the public. Boisjoly had failed to prevent the disaster but he had exercised his professional responsibilities as he saw them." (Harris et al. 2000: 4f.)

Under these presuppositions, the story seems to show us dramatically how important it is that, in accomplishing their responsibility, engineers stick to their professional knowledge and obligations and resist various influences which come from outside of engineering.

On the other hand, we cannot ignore the fact that this story has decisive weaknesses.

One of the problems of the story based on these presuppositions is that it is understandable only in hindsight. If we take seriously the actual situation in the past, we cannot easily presuppose that Boisjoly really understood the problem of the O-ring very well.

First, up to the teleconference on the eve of the Challenger launch, Boisjoly believed that the joint was an acceptable risk because of the redundancy of the second O-ring (Vaughan 1996: 187). Second, if he had really known the problem of O-ring, he could have demonstrated the correlation between temperature and resiliency in a much more definite and persuasive way, and above all not on the eve of the launch but much earlier. We cannot characterize someone's belief as a genuine knowledge, which turns out to be true afterwards, as long as it could not be persuasively demonstrated to be true when it was demanded. Third, it is only a groundless supposition that Boisjoly might have done more than he really did, as he himself "felt he had 'done everything he could'," and "it is also questionable that Boisjoly believed in the fatal consequence of the launch under the expected condition, as even Boisjoly acted as if he expected the mission to return" (Vaughan 1996: 380). In addition, even managers would have not allowed the launch, if there had been clear evidence of the possible danger. What is missing in the story is an understanding of the character of technological knowledge and judgment.

What is characteristic in engineers' activity is that engineers must judge and make decisions in uncertain situations in which no clear or definitive answer can be found in advance. In this sense, Boisjoly's judgment must be regarded as one possible judgment among others, and therefore the conflict concerning whether the launch could be approved or not is to be found between the engineers and managers and among the engineers themselves (Vaughan 1996: 324ff., 334, 356).

In addition to this problem, the story is also problematic in its narrative of the behavior of Boisjoly because the story ends only with admiration of Boisjoly's behavior and we can find no recommendations or suggestions as to how Boisjoly could have acted to prevent the accident. The story might indicate that even if engineers act ethically as engineers accidents cannot be avoided. In other words, it could suggest that it is possible to be assumed to be sufficiently ethical as the engineer in a self-contained way, independent of the ultimate results.

In this context, we can find an interesting episode in the report of the investigation board of Columbia. During the Columbia flight, and after finding out that the request for an image of the orbiter from some outside source had been cancelled by the managers, one engineer wrote an e-mail in which he emphasized that the damage by the debris could possibly bring about a very dangerous result, citing one of the mottoes of NASA, "If it is not safe, say, so". Considering the content of the e-mail, we can imagine that the engineer must have known very well what an engineer should do in such a situation. However, he did not send it. Instead he printed out and shared it with a colleague.

> "When asked why he did not send this e-mail, Rocha replied that he did not want to jump the chain of command." (Report 2003: 157)
> "Further, when asked by investigators why they were not more vocal about their concerns, Debris Assessment Team members opined that by raising contrary points of view about Shuttle mission safety, they would be singled out for possible ridicule by their peers and managers." (Report 2003: 169)

It seems that there was no "Boisjoly", at least Boisjoly à la story 1, in the case of Columbia. However, if a possible lesson of Challenger can be found in the point that even Boisjoly could not prevent an accident, it is understandable that engineers became skeptical about their chances of preventing accidents by trying to be more vocal and committing themselves to a kind of (near) whistle blowing action.

Of course this is not a logically necessary conclusion derived from the orthodox Challenger narrative. However, it cannot be denied that the possibility of drawing such a conclusion remains as long as, on the basis of this story, we have no

indication as to the question of what Boisjoly could have done to prevent a possible accident beyond what he actually did.

(2) Story 2: the normalization of deviance

If we leave the perspective in which hindsight is dominant and go back to the real situation in the past, when engineers were confronted with various uncertainties, we find a very different story. This revised story, which was originally written by a sociologist, Diane Vaughan, is so impressive and persuasive that many researchers use it to criticize the orthodox story (Vaughan 1996; Collins and Pinch 1998; Lynch and Kline 2000).

In focusing on the process of (social) construction of "acceptable risk", which plays a decisive role in engineers' judgment and decision, this revised story gives us an answer to the questions raised above, i.e. questions of why NASA continued to fly with a known O-ring erosion problem in the years before the Challenger launch, and why, on the eve of the Challenger launch, NASA managers decided that launching the mission in such cold temperatures was an acceptable risk, despite the concerns of their engineers.

First of all, we must confirm that there is no absolute certainty in the realm of engineering and that we can never objectively know the amount of risk. Larry Wear, an engineer at the Marshal Space Center, expressed this situation in the following way.

> "Any airplane designer, automobile designer, rocket designer would say that [O-ring] seals have to seal. They would all agree on that. But to what degree do they have to seal? There are no perfect, zero-leak seals. All seals leak some. [----] How much is acceptable? Well, that gets to be very subjective, as well as empirical." (Vaughan 1996: 115)

At least until the eve of the launch of the Challenger shuttle, engineers regarded the problems concerning the O-ring to be acceptable risk. For this interpretation of the O-ring's problem to be changed, there would have had to have been some decisive evidence. Boisjoly thought the apparent correlation between temperature and resiliency was sufficient evidence, but others did not think so. How should the dispute have been settled? Exactly in the way engineers and managers did on the eve of the launch.

> "Without hindsight to help them the engineers were simply doing the best expert job possible in an uncertain world. We are reminded that a risk-free technology is impossible and that assessing the working of a technology and the risks attached to it are always inescapable matters of human judgment." (Collins and Pinch 1998: 55)

In this way, we can find nothing special in the activities of engineers on the eve of the launch. Engineers and managers acted according to the established rules, just as in the case of a normal flight readiness review meeting. According to one

manager of NASA,"with all procedural systems in place, we had a failure" (Vaughan 1996: 347). At least in this sense, they did their best as usual but failed.

However, if what the engineers did on the eve of the launch can be considered the best action engineers can take, we again become perplexed by the conclusion of the story. Was it inevitable that the accident occurred? Was there no way to prevent it? And: Was there no lesson to be learnt from this story?

Perhaps the only lesson would be that the accident was inevitable in a complex technological system. While Vaughan's conclusion seems to be close to this pessimistic view, she tries to draw some lessons from her story. The lessons we should learn, however, can be drawn from the event which occurred on the eve of the launch, and from the preceding process of judgments and actions, in which the degree of acceptable risk concerning the O-ring was gradually increased. To explain this process, Vaughan proposed the term "normalization of deviance" (Vaughan 1996).

In many launches before that of Challenger, engineers had found various cases of erosion and blow-by of O-rings. However, what they found in these cases was not interpreted as an indicator of a safety problem but rather as evidence of acceptable risks, and as a result they step by step widened the range of acceptability. "The workgroup calculated and tested to find the limits and capabilities of joint performance. Each time, evidence initially interpreted as a deviance from expected performance was reinterpreted as within the bounds of acceptable risk" (Vaughan 1998: 120). Once such a process of normalization of deviance is begun and then gradually institutionalized, it is very difficult to stop this process of a "cultural construction of risk" (Vaughan 1998: 120). The only possible way to interfere with this process is to change the culture in which such a process is embedded and regarded as self evident, requiring a paradigm shift, such that anomalies that were neglected in the former paradigm become a focus (Vaughan 1996: 348,394).

This brings us close to the conclusions drawn by the Columbia accident investigation board, which stated NASA's history and culture should be considered the ultimate causes of the accident.

What then can we do to change the culture of an organization and prevent possible accidents?

It seems there is no special method immediately available. Changing an organizational structure and introducing new rules and guidelines would be possible measures; and these measures were taken within NASA after the Challenger accident. However, there is no guarantee that the realization of these measures will create a better situation, in which accidents would be prevented. On the contrary, there is even a possibility that we will introduce a new hazard, just as we often find in cases of design change. It is well known that any design change, no matter how seemingly benign or beneficial, has the potential to introduce a possibility for failure (Petroski 1994: 57)

> "Perhaps the most troubling irony of social control demonstrated by this case [structural change done by NASA after the Challenger accident] is that the rules themselves can have unintended effects on system complexity and, thus, mistake. The number of guidelines—and conformity to them—may increase risk by giving a false sense of security" (Vaughan 1996: 420).

This comment, which was written by Vaughan before the Columbia accident, was unfortunately verified by the Columbia accident.

3. Normal accidents and responsibility as civic virtue

The second, revised story of the Challenger accident seems to be much more realistic than the first one, but the conclusion derived from it seems to be much worse, or at least more pessimistic. Can we gain some lessons concerning engineering ethics from it? In an attempt to find a possible set of ethics which takes into account lessons derived from the second story seriously, I will consider several ideas proposed by two thinkers, Charles Perrow and John Ladd (Perrow 1999; Ladd 1991).

(1) Normal accidents

On the basis of the analysis of the accident of the nuclear plant at Three Mile Island, Charles Perrow proposed the term "normal accident" to characterize what happens with high-risk technologies. In highly complex technological organizations where factors are tightly connected, accidents occur in an unpredictable, inevitable and incomprehensible way.

These elements of unpredictability, inevitability and incomprehensibility are not a factual limit, which we can overcome with some new knowledge or technologies. Rather every effort to overcome the limit of these characters cannot but make an organization more complex and produce new possible dangers.

> "If interactive complexity and tight coupling—system characteristics—inevitably will produce an accident, I believe we are justified in calling it a *normal accident*, or a *system accident*. The odd term *normal accident* is meant to signal that, given the system characteristics, multiple and unexpected interactions of failure are inevitable" (Perrow 1999: 5)

The term "normal accident" is very insightful, as it indicates that in high-risk technologies the normal processes of engineers' activities at workplace are to be considered processes for producing products and, simultaneously as processes that produce new hazards. If engineers want to avoid committing themselves to such processes and take a conservative position as far as possible, the only possible way for them to work would be to restrict themselves to working within a laboratory. In this context, Vaughan cites an interesting expression "engineering purist", which is used by Marshall's engineers to characterize an engineer, who works only in a laboratory, does not have to make decisions and can take the most conservative position in the world (Vaughan 1996: 88). In contrast to this "engineering purist", every engineer who works and makes decisions in a real workplace, in which not only "purely" technical problems but various kinds of conditions such as cost and schedule must be taken into consideration, can never take the most conservative position.

If we take these circumstances seriously, we cannot but change our view of the meaning of the everyday activities of engineers. For example, if we follow this normal accident view, every decision process about a certain acceptable risk must also be regarded as a process at the same time producing another possible risk, and therefore previous success cannot be used as a justification for accepting increased risk. Perhaps this sounds a little extreme: but we can find this kind of warning in the statements of working engineers. Petroski emphasizes that if engineers design new things past success is no guarantee of the success of new design and cites the following statements of engineers. "Engineers should be slightly paranoiac during the design stage". "I look at everything and try to imagine disaster. I am always scared. Imagination and fear are among the best engineering tools for preventing tragedy" (Petroski 1994: 3, 31). If engineers could continue to take this kind of view in every step of their work, the process of normalization of deviance would not remain invisible but would inevitably come to the fore.

What is important here is that this kind of change of attitudes cannot be realized by changing explicit rules or institutional structures, as the main point is that these changes always have the potential to produce new risks. It is remarkable that the recommendations made by the Columbia accident investigation board focus on this point.

For example, in the report the normal accident theory is used to analyze the causes of the accident. The report indicates the need for a change of culture within NASA and makes the following proposals.

> "The [Space Shuttle] Program must also <u>remain sensitive</u> to the fact that <u>despite its best intention</u>, managers, engineers, safety professional,

> and other employees, can when confronted with extraordinary demands, act in counterproductive ways" (Report 2003: 181).
>
> "Organizations that deal with high-risk operations must always have a healthy fear of failure—operations must be proved safe, rather than the other way around." (Report 2003: 190)

These sentences suggest clearly where we should search for resources to change the culture in question. Surely not in the ethics in the narrow sense of the word, as "best intentions" people might have cannot contribute to preventing failures. Rather "sensitivity" to possible accidents and "a healthy fear of failure" must play a decisive role.

What kind of ethics would we have, if we take these indications seriously?

(2) Civic virtue

On the basis of the analysis of one typical case of a normal accident, the catastrophy at the chemical factory Union Carbide at Bhopal in India, John Ladd attempts to identify the ethical dimension indicated by cases of normal accidents by proposing the interesting concept of "civic virtue".

Ladd introduces a difference concerning the concept of responsibility. One is a narrow, legal and negative concept of responsibility, which is also characterized as job-responsibility or task-responsibility. If someone does not fulfill this responsibility, he or she will be blamed. In this understanding, the concept of responsibility is used exclusively, and the concept of non-responsibility plays as important a role as the concept of responsibility, and the question of *who* is responsible and *who is not* is important in this context. "We hear claims of responsibility voiced in hearing 'It's my job, not his' as well as disclaimers of responsibility in hearing 'It's his job, not mine'" (Ladd 1991:81).

In contrast to this kind of concept, the second concept of responsibility is characterized as broad, moral and positive.

According to this concept, even if someone does not fulfill a responsibility, it is not necessary that he or she will be blamed. In other words, if someone is responsible for something in this sense, it does not exclude others from also being responsible. In this sense, the "collective responsibility" of a large part of the population for the same thing is possible (Ladd 1991, 81). This moral responsibility is "something positively good, that is, something to be sought after" and "something that good people are ready and willing to acknowledge and to embrace" (Ladd 1991: 82).

Ladd calls this kind of responsibility "civic virtue". It is characterized, firstly, as moral virtue, because it contains as an essential factor an attitude of concern for the welfare of others, i.e. humanity. Secondly, as this attitude of caring and regarding of others is a virtue everyone should have when exercising relationships with others, this responsibility is characterized as civic.

> "Our attitude towards whistle blowing illustrate how far we have gone in turning our values upside down: the concern for safety, which should motivate all of us, has been relegated to the private realm of heroes, troublemakers and nuts. Our society assumes that it is a matter of individual choice (and risk) to decide whether or not to call attention to hazards and risks instead of being, as it should be, a duty incumbent on all citizens as responsible members of society.
> This is where virtue comes in, or what in the present context I shall call *civic virtue*. Civic virtue is a virtue required of all citizens. It is not just something optional—for saints and heroes." (Ladd 1991: 90)

To this last sentence we could add the word 'engineers.' According to this view, to prevent hazards and risks is not the special responsibility of engineers as professionals but rather the universal responsibility of all citizens.

If we relate the indications derived from the concept of normal accident in the last section to this discussion, we will become able to add some content to the concept of the responsibility as civic virtue.

"Sensitivity" to a possible danger and "a healthy fear of a failure", which can be regarded as essential factors constituting a culture of safety, must be a central feature of civic virtue, a feature which can contribute to the prevention of possible accidents.

As long as we remain in the dimension of negative responsibility, it is difficult to identify someone who is to be blamed in the case of normal accidents, but it is also unhelpful to do so, as the replacement of the individuals to be blamed would not necessarily change the culture of the organization. If we look at the situation from the standpoint of positive responsibility, we can find many irresponsible acts, such as lack of concern, negligence in the face of signs of a hazard and so on, which are rooted in a general culture, exactly as in the cases of Challenger and Columbia. In this way, from the view of civic virtue, we can take into account the collective responsibility of an organization and indicate a need to change its culture to prevent accidents. In this sense, the concept of civic virtue is to be understood as a virtue belonging to an individual and as a virtue belonging to an organization.

In addition to it, as Vaughan and the report of the Columbia accident investigation board emphasize, the organizational culture of NASA is constrained and constituted by external political and economical factors decided by the USA Congress and the White House. From the point of view of civic virtue, we could extend the scope of collective responsibility to the people belonging to these organizations, as these people, as responsible citizens, cannot evade responsibilities, just as the NASA administrators, middle level managers and engineers cannot evade responsibilities. The concept of civic virtue must be ascribed to every responsible and related person. In this sense, if one wants to promote a hazard aware environment, where people work, act and govern ethically, it is necessary to cultivate professional responsibility but more importantly, civic virtue must be nurtured within the organization and society. Such civic virtue is rooted in a capacity to respond to and care for others and thus constitutes a fundamental dimension of ethics.

4. Conclusion

Considering all of these discussions, what lesson can we learn for engineering ethics?

Firstly, as already indicated, engineering ethics is usually characterized as professional ethics. This kind of ethics might be very helpful for producing honest, loyal and "responsible" engineers who can solve the various problems they confront in their work place and accomplish their work as engineers. However, as long as engineering ethics remains in the dimension of professional ethics, based around the actions of the engineer, engineering ethics will fail the ultimate goal of being a means by which engineers can be made to think about, and take responsibility for, preventing possible disasters that could result from their everyday normal practices. To fulfill this role engineering ethics needs to encompass factors which are rooted in a much more fundamental dimension than that of professional ethics.

Secondly, "engineering ethics" is commonly classified as ethics on the micro level in contrast to the "ethics of technology", in which philosophical and political problems concerning the relationship between technology and society on the macro level is discussed. Surely we cannot confuse the different levels of discussions. However, when it comes to preventing disastrous "normal" accidents, and when causes of normal accidents are rooted in organizational culture, which is inseparably connected with macro level factors, we cannot leave the discussion of engineering ethics within the micro and individual dimension but must extend its discussion and connect it to the discussions taking place on the macro level. The concepts of "culture" and "civic virtue" can be used to mediate between the two levels of discussions thus extending and making more fruitful the field of "engineering ethics ".

References

Collins, Harry and Trevor Pinch. 1998. *The Golem at Large: what you should know about technology*. Cambridge: Cambridge University Press.

Davis, Michael. 1998. *Thinking Like an Engineers, Studies in the Ethics of a Profession*. Oxford: Oxford University Press.

Harris, Charles, Michael Pritchard and Michael Rabins. 2000. *Engineering Ethics, Concepts and Cases*, second edition. Belmont, CA: Wadsworth.

Johnson, Deborah G. 2001. *Computer Ethics*. Englewood Cliffs: Prentice Hall.

Ladd, John. 1991. Bhopal: An Essay on Moral Responsibility and Civic Virtue. *Journal of Social Philosophy* 22(1): 73-91.

Lynch, William and Ronald Kline. 2000. Engineering Practice and Engineering Ethics. *Science, Technology and Human Values* 25(2): 195-225.

Murata, Junichi. 2003a. Creativity of Technology: An Origin of Modernity?. In *Modernity and Technology*, edited by Thomas Misa, Philip Brey, and Andrew Feenberg. Cambridge MA: The MIT Press.

Murata, Junichi. 2003b. Philosophy of Technology, and/or, Redefining Philosophy. *UTCP Bulletin* 1. University of Tokyo, Center for Philosophy: 5-14.

Murata, Junichi. 2003. Technology and Ethics—Pragmatism and the Philosophy of Technology. *The Proceedings for the UTCP International Symposium on Pragmatism and the Philosophy of Technology*, Volume 2: 60-70.

Perrow, Charles. 1999. *Normal Accidents, Living with High-Risk Technologies*. Princeton NJ: Princeton University Press.

Petroski, Henry. 1994. *Design Paradigms, Case Histories of Error and Judgment in Engineering*. Cambridge: Cambridge University Press.

Report. 2003. *Columbia Accident Investigation Board, Report* Volume 1, August 2003. Washington D.C.: Government Printing Office.

Tenner, Edward. 1996. *Why Things Bite Back, Technology and the Revenge Effect*. London: Fourth Estate.

Vaughan, Diane. 1996. *The Challenger Launch Decision, Risky Technology, Culture, and Deviance at NASA*. Chicago: The University of Chicago Press.

Safe Design

Sven Ove Hansson
Department of Philosophy and the History of Technology
Royal Institute of Technology, Stockholm
soh@infra.kth.se

Abstract: Safety is an essential ethical requirement in engineering design. Strategies for safe design are used not only to reduce estimated probabilities of injuries but also to cope with hazards and eventualities that cannot be assigned meaningful probabilities. The notion of safe design has important ethical dimensions, such as that of determining the responsibility that a designer has for future uses (and misuses) of the designed object.

Keywords: safety, risk, safe design, safety barrier, ethics.

1. Safety – An ethical issue in design

In the small literature that is available on the ethics of engineering design, there is consensus that safety is an essential ethical requirement. It is generally agreed that designers have an ethical responsibility to make constructions that are safe in future use. However, it is far from clear how far this responsibility extends. It needs to be specified in at least two respects.

The first of these consists in answering the question "*safe against what?*" Safety is concerned with avoiding certain classes of events that it is morally right to avoid. In engineering design, safety always includes safety against unintended human death or injuries that occur as a result of the intended use of the designed object. But does it include the avoidance of accidents in foreseeable but unintended uses of the object? Does it include protection against malevolent use of the object by criminals or terrorists? (Kemper 2004) The prevention of long-term health effects? The prevention of damage to the environment?

We can use the design of bridges as an example of this problem. Designers of bridges are normally held responsible for the structural reliability of their constructions. If a bridge collapses, then we hold the engineers who designed it responsible. However, there are other types of safety issues in connection with bridges. Accidents happen when people climb and walk on arches, dive from the bridge, or throw objects on ships or vehicles passing below the bridge. Dark and inaccessible parts of bridges can be used for criminal activities. Some people commit suicide by jumping from bridges. Most of these issues are not traditionally taken to be the responsibility of bridge constructors. (van Gorp 2005, pp.104-110) Should the concept of safe design be so wide that it covers these and other potential negative events in addition to the traditional issues of structural reliability?

It can be argued in favour of a wide definition of the designer's responsibility that what she does has a lasting influence on safety. The designer can often solve safety problems that are virtually impossible for future users to solve. However, against this it can be argued that the designer is not in a position to solve all problems that may arise from future uses. It is impossible to predict all future uses and misuses of a product. How can the designer be responsible for future events that she has no means to foresee?

The other aspect of safety that needs to be specified is *what it means to be safe against something*. This is the subject of the present contribution. I will approach it by studying some major practices in engineering design.

2. Practices in Safe Design

There are many treatments of safe design in particular fields of engineering, but I am not aware of any fully general account of principles for safe design. However, the following four design principles are in general use in many fields of engineering. They can therefore be taken as representative of the engineering practices of safe design:

1. *Inherently safe design*. A recommended first step in safety engineering is to minimize the inherent dangers in the process as far as possible. This means that potential hazards are excluded rather than just enclosed or otherwise coped with. Hence, dangerous substances or reactions are replaced by less dangerous ones, and this is preferred to using the dangerous substances in an encapsulated process. Fireproof materials are used instead of inflammable ones, and this is considered superior to using flammable materials but keeping temperatures low. For similar reasons, performing a reaction at low temperature and pressure is considered superior to performing it at high temperature and pressure in a vessel constructed for these conditions.

2. *Safety factors*. Constructions should be strong enough to resist loads and disturbances exceeding those that are intended. A common way to obtain such safety reserves is to employ explicitly chosen, numerical safety factors. Hence, if a safety factor of 2 is employed when building a bridge, then the bridge is calculated to resist twice the maximal load to which it will in practice be exposed.

3. *Negative feedback*. Negative feedback mechanisms are introduced to achieve a self-shutdown in case of device failure or when the operator loses control. Two classical examples are the safety-valve that lets out steam when the pressure becomes too high in a steam-boiler and the dead man’s handle that stops the train when the driver falls asleep. One of the most important safety measures in the nuclear industry is to ensure that reactors close down automatically in critical situations.

4. *Multiple independent safety barriers*. Safety barriers are arranged in chains. The aim is to make each barrier independent of its predecessors so that if the first fails, then the second is still intact, etc. Typically the first barriers are measures to prevent an accident, after which follow barriers that limit the consequences of an accident, and finally rescue services as the last resort. One of the major lessons from the Titanic disaster is that an improvement of the early barriers (in this case: a hull divided into watertight compartments) is no excuse for reducing the later barriers (in this case: lifeboats).

Safety engineering includes many more principles and practices than the four mentioned above. Education of operators, maintenance of equipment and installations, and incidence reporting are examples of safety practices of general importance. However, I believe that the four mentioned above cover at least a large part of the practices that are central in engineering design.

3. SAFETY, RISK, AND UNCERTAINTY

Is there a common notion of safety underlying the four general safety practices outlined in the previous section? One obvious answer could be that safety is understood in this context as the antonym of risk, so that a design is safe to the extent that it reduces risk. In probabilistic risk analysis (PRA; also called probabilistic safety analysis, PSA), risk is defined in exact numerical terms. Therefore, safe design could tentatively be defined as design that reduces or minimizes risk in the standard sense of this term, as it is used in PRA. In what follows I will show that this is not a workable definition of safe design. To see this, we need to introduce the decision-theoretical distinction between risk and uncertainty.

In decision theory, "risk" and "uncertainty" are the two major categories of lack of knowledge. In decision-making under risk, the probabilities of possible outcomes are known, whereas in decision-making under uncertainty, probabilities are either unknown or only known with insufficient precision. Hence, decisions at the roulette table are decisions under risk, whereas a choice between two dinner parties is a decision under uncertainty. Uncertainty also covers the cases in which the possible outcomes, not only their probabilities, are unknown. (Hansson 1996)

Few if any decisions in actual life are based on probabilities that are known with certainty. Strictly speaking, the only clear-cut cases of "risk" (known probabilities) seem to be idealized textbook cases that refer to devices such as dice, coins, or roulette wheels that are supposedly known with certainty to be fair. More typical real-life cases are characterized by uncertainty that does not, primarily, come with exact probabilities. Hence, almost all decisions are decisions "under uncertainty". To the extent that we make decisions "under risk", this does not mean that these decisions are made under conditions of completely known probabilities. Rather, it means that we have chosen to simplify our description of these decision problems by treating them as cases of known probabilities.

This ubiquity of uncertainty applies also in engineering design. An engineer performing a complex design task has to take into account a large number of hazards and eventualities. Some of these eventualities can be treated in terms of probabilities; the failure rates of some components may for instance be reasonably well-known from previous experiences. However, even when we have a good experience-based estimate of a failure rate, some uncertainty remains about the correctness of this estimate and in particular about its applicability in the context to which we apply it. In addition, in every design process there are uncertainties for which we do not have good or even meaningful probability estimates. This includes the ways in which humans will interact with new constructions. As one example of this, users sometimes "compensate" for improved technical safety by more risk-taking behaviour. Drivers are known to have driven faster or delayed braking when driving cars with better brakes. (Rothengatter 2002) It is not in practice possible to assign meaningful numerical probabilities to these and other human reactions to new and untested designs. It is also difficult to determine adequate probabilities for unexpected failures in new materials and constructions or in complex new software. We can never escape the uncertainty that refers to the eventuality of new types of failures that we have not been able to foresee.

Of course, whereas reducing risk is obviously desirable, the same may not be said about the reduction of uncertainty. Strictly interpreted, uncertainty reduction is an epistemic goal rather than a practical one. However, by reducing uncertainty we place ourselves in a situation in which we can make more well-informed practical decisions, e.g. about risk reduction. In the choice between decision alternatives that differ in their degrees of uncertainty about possible dangers, by choosing an alternative with low uncertainty we ensure that risks are within stricter bounds than if we choose an alternative with greater uncertainty in this respect.

In summary, engineering design always has to take into account *both* uncertainties that can be meaningfully expressed in probabilistic terms *and* eventualities for which this is not possible. The former are no less ethically relevant than the latter. In the next two sections, I will discuss the implications of uncertainty for two of the safe design strategies mentioned above, namely safety factors and multiple safety barriers.

4. Safety Factors

Probably, humans have made use of safety reserves since the origin of our species. They have added extra strength to their houses, tools, and other constructions in order to be on the safe side. However, the use of numerical factors for dimensioning safety reserves seems to be of relatively recent origin, probably the latter half of the 19th century. The earliest usage of the term recorded in the Oxford English Dictionary is from WJM Rankine's book *A manual of applied mechanics* from 1858. In the 1860s, the German railroad engineer A. Wohler recommended a factor of 2 for tension. (Randall 1976) The use of safety factors is now since long well established in structural mechanics and in its many applications in different engineering disciplines. Elaborate systems of safety factors have been developed, and specified in norms and standards.

A safety factor is typically intended to protect against a particular integrity-threatening mechanism, and different safety factors can be used against different such mechanisms. Hence one safety factor may be required for resistance to plastic deformation and another for fatigue resistance. As already indicated, a safety factor is most commonly expressed as the ratio between a measure of the maximal load not leading to the specified type of failure and a corresponding measure of the applied load. In some cases it may instead be expressed as the ratio between the estimated design life and the actual service life.

In some applications safety margins are used instead of safety factors. A safety margin differs from a safety factor in being additive rather than multiplicative. In order to keep airplanes sufficiently apart in the air a safety margin in the form of a minimal distance is used. Safety margins are also used in structural engineering, for instance in geotechnical calculations of embankment reliability. (Duncan 2000)

According to standard accounts of structural mechanics, safety factors are intended to compensate for five major categories of sources of failure:

1) higher loads than those foreseen,
2) worse properties of the material than foreseen,
3) imperfect theory of the failure mechanism in question,
4) possibly unknown failure mechanisms, and
5) human error (e.g. in design).

(Knoll 1976. Moses 1997.)

The first two of these refer to the variability of loads and material properties. Such variabilities can often be expressed in terms of probability distributions. However, when it comes to the extreme ends of the distributions, lack of statistical information can make precise probabilistic analysis impossible. Let us consider the variability of the properties of materials. Experimental data on material properties are often insufficient for making a distinction between e.g. gamma and lognormal distributions, a problem called *distribution arbitrariness*. (Ditlevsen 1994) This has little effect on the central part of these distributions, but in the distribution tails the differences can become very large. This is a major reason why safety factors are often used as design guidance instead of probabilities, although the purpose is to protect against failure types that one would, theoretically, prefer to analyze in probabilistic terms.

> Theoretically, design by using structural system reliability is much more reasonable than that based on the safety factor. However, because of the lack of statistical data from the strength of materials used and the applied loads, design concepts based on the safety factor will still dominate for a period. (Zhu 1993)

The last three of the five items on the list of what safety factors should protect against all refer essentially to errors in our theory and in our application of it. They are therefore clear examples of uncertainties that are not easily amenable to probabilistic treatment. In other words: The eventuality of errors in our calculations or their underpinnings is an important reason to apply safety factors. This is an uncertainty that is not reducible to probabilities that we can determine and introduce into our calculations. It is for instance difficult to see how a calculation could be accurately adjusted to compensate self-referentially for the possibility that it may itself be wrong. However, these difficulties do not make these sources of failures less important from an ethical point of view. Safety factors are used to deal both with those failures that can be accounted for in probabilistic terms and those that cannot.

5. Safety Barriers

Some of the best examples of the use of multiple safety barriers can be found in nuclear waste management. The proposed subterranean nuclear waste repositories all contain multiple barriers. We can take the current Swedish nuclear waste project as an example. The waste will be put in a copper canister that is constructed to resist the foreseeable challenges. The canister is surrounded by a layer of bentonite clay that protects the canister against small movements in the rock and "acts as a filter in the unlikely event that any radionuclides should escape from a canister". This whole construction is placed in deep rock, in a geological formation that has been selected to minimize transportation to the surface of any possible leakage of radionuclides. The whole system of barriers is constructed to have a high degree of redundancy, so that if one the barriers fails the remaining ones will suffice. With usual PRA standards, the whole series of barriers would not be necessary. Nevertheless, sensible reasons can be given for this approach, namely reasons that refer to uncertainty. Perhaps the copper canister will fail for some unknown reason not included in the calculations. Then, hopefully, the radionuclides will stay in the bentonite, etc. In this particular case, redundancy can also be seen as a means to meet public scepticism and opposition (although it is not self-evident that redundant safety barriers will make the public feel safer).

For another example, we can again consider what is possibly the most well-known example of technological failure in modern history, the Titanic that sank with 1500 persons in April 1912. It was built with a double-bottomed hull that was divided into sixteen compartments, constructed to be watertight. Four of these could be filled with water without danger. Therefore, the ship was believed to be unsinkable, and consequently it was equipped with lifeboats only for about half of the persons onboard.

We now know that the Titanic was far from unsinkable. But let us consider a hypothetical scenario. Suppose that tomorrow a ship-builder comes up with a convincing plan for an unsinkable boat. A probabilistic risk analysis shows that the probability of the ship sinking is incredibly low. Based on the PRA, a risk-benefit analysis has been performed. It shows that the cost of life-boats would be economically indefensible. The expected cost per life saved by the life-boats is above 1000 million dollars, a sum that can evidently be more efficiently used to save lives elsewhere. The risk-benefit analysis therefore clearly shows us that the ship should not have any lifeboats.

How should the naval engineer respond to this proposal? Should she accept the verdict of the economic analysis and exclude lifeboats from the design? My proposal is that a good engineer should not act on the risk-benefit analyst's advice in a case like this. The reason for this is obvious from what has already been said: The calculations may possibly be wrong, and if they are, then the outcome may be disastrous. Therefore, the additional safety barrier in the form of lifeboats (and evacuation routines and all the rest) should not be excluded, in spite of the probability estimates showing them to be uncalled for.

6. Conclusion

Many of the most ethically important safety issues in engineering design refer to hazards that cannot be assigned meaningful probability estimates. It is appropriate that at least two of the most important strategies for safety in engineering design, namely safety factors and multiple safety barriers, deal not only with risk (in the standard, probabilistic sense of the term) but also with uncertainty.

Currently there is a trend in several fields of engineering design towards increased use of probabilistic risk analysis (PRA). This trend may be a mixed blessing since it can lead to a one-sided focus on those dangers that can be assigned meaningful probability estimates. PRA is an important design tool, but it is not the final arbitrator of safe design since it does not deal adequately with issues of uncertainty. Design practices such as safety factors and multiple barriers are indispensable in the design process, and so is ethical reflection and argumentation on issues of safety. Probability calculations can often support, but never supplant, the engineer's ethically responsible judgment.

References

Clausen, Jonas, Sven Ove Hansson and Fred Nilsson, "Generalizing the Safety Factor Approach", *Journal of Reliability and Engineering System Safety,* in press.

Ditlevsen, O. 1994. "Distribution arbitrariness in structural reliability" in Schuëller, G. Shinozuka, M. and Yao, J. (1994) *Proc. of ICOSSAR'93: Structural Safety & Reliability* 1241-1247.

Duncan, J.M. 2000. "Factors of safety and reliability in geotechnical engineering". *Journal of Geotechnical and Geoenvironmental Engineering* 126:307-316.

Hansson, Sven Ove. 1996 "Decision-Making Under Great Uncertainty", *Philosophy of the Social Sciences* 26:369-386.

Kemper, Bart. 2004. "Evil Intent and Design Responsibility" *Science and Engineering Ethics* 10(2): 303-309.

Knoll, F. 1976. “Commentary on the basic philosophy and recent development for safety margins”, *Canadian Journal of Civil Engineering.* 3:409-416.

Lloyd.Peter and Jerry Busby.2003. “Things That Went Well—No Serious Injuries or Deaths”: Ethical Reasoning in a Normal Engineering design Process” *Science and Engineering Ethics* 9:503-516.

Martin, Mike W. and Roland Schinzinger. 2005. *Ethics in engineering*, 4th ed., Boston: McGraw-Hill, 2005.

Moses, F. 1997. “Problems and prospects of reliability-based optimisation”, *Engineering Structures* 19:293-301.

Palm, Elin and Sven Hansson, “The Case for Ethical Technology Assessment (eTA), *Technological Forecasting and Social Change,* in press.

Randall, F.A. 1976. “The safety factor of structures in history”, *Professional Safety* January:12-28.

Rothengatter, Talib. 2002 “Drivers’ illusions – no more risk”, *Transportation Research*, part F, 5:249-258.

van de Poel, Ibo. 2001 “Investigating Ethical Issues in Engineering Design” *Science and Engineering Ethics* 7: 429-446.

van Gorp, Anke. 2005. *Ethical issues in engineering desing; Safety and sustainability*, PhD thesis, Delft University 2005.

Zhu, T.L. 1993. “A reliability-based safety factor for aircraft composite structures”, *Computers & Structures* 48:745-748.

Social Norms in Artefact Use: Proper Functions and Action Theory

Marcel Scheele

Abstract: The use of artefacts by human agents is subject to human standards or norms of conduct. Many of those norms are provided by the social context in which artefacts are used. Others are provided by the proper functions of the artefacts. This article argues for a general framework in which norms that are provided by proper functions are related to norms provided by the (more general) social context of use. Departing from the concept, developed by Joseph Raz, of "exclusionary reasons" it is argued that proper functions provide "institutional reasons" for use. Proper use of artefacts (use according to the proper function) is embedded in the normative structures of social institutions. These social normative structures are complementary to traditional norms of practical rationality and are a kind of second-order reasons: exclusionary reasons. It is argued that proper functions of artefacts provide institutional reasons, which are up to a certain extent similar to exclusionary reasons. The most notable difference concerns the fact that proper functions not so much exclude other types of use, but rather place that use (and the user) in particular social structures with particular rights and obligations. An institutional reason not only gives a reason for action, it also provides reasons for evaluating actions according to such reasons positively (and other negatively). The upshot of the analysis is that it provides an additional tool for understanding and evaluating the use of artefacts.

Keywords: proper function; normativity; exclusionary reason; institution

1. Introduction

This article is about the use of artefacts understood from an action theoretic perspective. Use is a kind of intentional action and is as such guided by norms. Just as with every-day action, people sometimes abide by those norms and sometimes they don't. Whatever the content of these norms are, they can help *understand* actions of people by making their reasons for action transparent. Norms are also used in the *evaluation* of action by making deliberated judgements possible about these actions. These can be used for the assignment of responsibility when needed.

In this article I argue that in the case of artefact use a special kind of norm plays a role that has had little or no attention up to now, certainly not in this form. This norm is not the only norm guiding artefact use; it is additional to the usual types of norms that are said to guide action. This is a kind of second-order (or higher-order) norm that structures the decision-space of action, rather than being directly part of the decision space. It is inspired by Joseph Raz' idea of *exclusionary reasons*. These are reasons that are derived from second-order norms (or higher-order norms). The reasons and norms analysed in this article share some features with exclusionary reasons, but are not completely identical to them. I call the reasons that are derived from the relevant norms *institutional reasons*. The reason for this terminology will become clear below.

As I said, artefact use is a kind of intentional action and can be analysed with the tools of action theory. Here we need to analyse "*to use*", rather than "*to do*". Using something can involve several different types of actions and of norms involved. What 'using' involves becomes clear from the following example. I can use a standard electric drill in a number of ways and these different uses can have a number of different results. I can use it for drilling a hole in a normal wall. I can use it to drill a hole in a concrete wall. But I can also use it to place it on a stack of papers to prevent them from being blown away. In fact, the number of uses I can put the drill to is unlimited. The results of these possible actions can be of several kinds. My action can succeed or fail. In both cases my action can have (only) positive effects or (only) negative effects, such as damage or injury. Such effects can be intended or unintended side effects. In all cases there are different ways in which the action can be evaluated. I can evaluate the actions and intentions of the user; I can evaluate the outcome; or a combination of these.

The outcome of the evaluation depends on the norms I use in evaluation. Relevant for these norms is the proper function of the artefact, what the artefact *is for*. The notion of 'proper function' thus plays an important role in the analysis below. This article gives a philosophical analysis of relevant norms of use by combining the functions of artefacts with further action theoretical norms. An action theoretical analysis of artefact use and the norms applicable to this should take functions into account. An account of use should help the interpretation and evaluation of use by providing a normative framework that gives criteria for interpretation and evaluation. In action theory the traditional standard normative framework is a framework of rationality: rational norms are used as constraints for action.

It has been argued in the past that rational norms are not sufficient to interpret action in general, because rational norms underdetermine action. Social norms should be added in order to create a complete interpretative and evaluative framework. Elsewhere I have argued that the proper functions of artefacts also should be understood as involving social norms, because the standard features that are said to determine the proper function underdetermine it (Scheele 2005a; 2006). The relation between the social norms pertaining to artefact functions and the social norms pertaining to artefact use will be investigated in this article.

The strategy is as follows. First I show how we should understand the proper function of artefacts when taking the social context of use into account (section 2). Then I describe how Raz conceives of the possibility of interpreting (certain) social norms as secondary reasons, supplementing rational norms of action (section 3). Then I argue that social norms pertaining to proper functions play a similar, but not identical role in the analysis of use. A somewhat different notion of secondary reasons needs to be introduced to understand the role of social norms in artefact use. I call these 'institutional reasons' (section 4).

2. The social factors partly determining the proper function

Artefact use can be seen as a category of intentional action; we use artefacts to achieve our goals. The physical capacities of these objects enable us to realise these goals. Each of the capacities that enable a possible use is called a *system function* and use according to such a function is called *effective use* or *rational use*. This type of use can be analysed within most standard action theoretical frameworks that involve means-end rationality. Effective use can be contrasted with use that is according to the *proper function* of an artefact and is called 'proper use'. The proper function denotes that *which the artefact is for*, denoting a privileged way of using the artefact.[1] This latter type of use stands central in

[1] This terminology follows Preston in her (Preston 1998). It is standard to read this notion in an ethically neutral sense. So it can be the proper function of a gun to shoot people with, although we may reject this use on other grounds.

this article because the proper function of artefacts provides the kind of norms I am interested in.

Proper functions differ from system functions due to their normative import. A normal functioning artefact has many system functions and one of them is the proper function.[2] In normal circumstances, therefore, there is no material distinction between a system function and a proper function. However, there *is* some distinction, which is most clearly seen when an artefact loses a function. If this function loss is of a mere system function, nothing else, besides the physical loss happens: the artefact simply can't be used for something it could be before. But if an artefact loses its proper function we say it *malfunctions*, which involves a normative judgement in addition.

A car, for example, that cannot move anymore, is said to malfunction, because its *proper* function is to ride and transport people and it has lost that particular capacity. However, if I put a cleaner engine in the car, the car loses its *system* function to help rapidly advance global heating, but doing that wasn't its proper function in the first place, so we cannot say it now malfunctions on that account. Adequately accounting for proper functions and for malfunctioning is an important challenge in function theory.

Philosophical attention for artefact functions is relatively recent and is only in the process of being separated from the notion of biological function, a notion that has been extensively researched in philosophy.[3] There are some similarities between biological functions and artefact functions. Restricting our attention to proper functions there are two central similarities. In the first place, biological function ascription, as well as artefact function ascription, takes the physical capacities of the object into account, although a qualification is added for the possibility of malfunction. In the second place the function, even in the absence of the right physical capacity, is justified in terms of the causal history of the item. In the case of biological functions the relevant causal history is determined by the item's *evolutionary history*. A modern classic is Karen Neander's definition:

[2] Although an artefact may have several proper functions simultaneously I disregard this most of the time for the sake of simplicity

[3] A useful anthology of function theories is (Buller 1999). A survey of function theories that also pays attention to artefact functions is (Perlman 2004).

> 'It is the/a proper function of an item (*X*) of an organism (*O*) to do that which items of *X*'s type did to contribute to the inclusive fitness of *O*'s ancestors, and which caused the genotype, of which *X* is the phenotypic expression, to be selected by natural selection.' (Neander 1991: 174)

This definition uses a historical notion, natural selection. A proper function of an item is a function to do something, namely that what in the past was done by items of that type, which helped make the ancestors of the organism more fit. By definition, this is the 'reason' that the item was selected by natural selection and hence justifies the function ascription.

In the case of artefacts the *design history* is generally conceived of as the central determining factor, as opposed to the selection history. This means that the function the designer (and/or manufacturer) intended for the artefact determines its proper function. Sometimes, in addition to the designer's intentions a kind of analogue to evolutionary history is introduced, in terms of the market mechanism which determines the success of the product (Preston 1998). So, although biological and artefact functions both refer to a causal history, the type of causal history referred to differs (cf. Millikan 1999).

Another difference, which is more important for our purposes, is the way in which the proper function is related to *using* the artefact. This difference is directly relevant to our discussion, because now we leave the domain of 'function theory' and enter that of 'action theory' and thus the study of norms for action. In general, although all organs may be *useful* for the creature that has them, the creature does not *use* all of them. With biological organs the proper function refers to the contribution to fitness an item has, whereas with artefacts the proper function refers to the proper use of the artefact, in combination with the way the artefact is supposed to work. We can say that in function theory the central analysandum for artefact functions differs from that for biological functions. Artefact functions directly involve intentional action.

It may be noted here that the distinction between "being used for" and "being useful for" does not coincide with the distinction between biological organs and artefacts. Biological organs *can* be used (by their owner) whereas artefacts, when functioning properly, are not always used intentionally (or consciously). Agents often *do* use their hands. And agents also generally *don't* use the CPU of their computer consciously for a specific goal. Also think of artificial hearts in this respect. These examples should be analysed carefully with respect to the question whether some item is merely useful or also (intentionally) used by an agent. "Being used" and "merely being useful for" certainly has vague boundaries. For purposes of this article I need not analyse these cases in detail and I concentrate on the cases for which proper use *is* clearly associated with the proper function (cf. Houkes and Vermaas 2004; Scheele 2005a).

This difference between artefacts and biological organs should be taken into account in the analysis of artefact functions and use. The analysis of artefact functions should take this aspect of *use* into account. I have argued elsewhere that proper functions of artefacts are partly determined in terms of the social environment in which the artefact is used. The social environment partly provides the norms of use and hence partly determines the proper use. In brief the point is as follows.[4] Although in most cases the proper functions of artefacts are determined by the designer or manufacturer this is not necessarily the case. An artefact that is used by everyone in a way alternative to the intentions of the designer will very soon change its proper function, due to the fact that this new use will become generally accepted. The reason for this is, roughly, that the socially accepted use will have changed, i.e. the relevant social norms will have changed and thus have overruled the original function ascription. This argument can be made general and shows that the original function determination also is, in part, socially determined: the designer and/or manufacturer has been assigned authority for this determination, which is a social category. The conclusion is that proper functions are partly constituted by the social acceptance of the use by a relevant community of users.[5]

However, invoking reference to 'social facts' or 'social acceptance' is still vague and does not make clear the (social) normative import that proper functions have. For the purposes of this article we are interested in this normative import. The relevant social aspects can be understood in terms of the *institutionalisation* of artefact use. The notion of 'institution' is generally used in the social sciences to indicate *enduring social structures*. The tendency is to treat it as a social scientific equivalent of 'substance'. It is used in many different disciplines and in

[4] For the details of the analysis and the arguments (cf. Scheele 2005a; 2006).

[5] This implies that identical artefacts(with regard to their physical structure as well as their production history) can have different proper functions in different social groups. These different groups thus have different norms concerning the use of that artefact.

various ways. From an action theoretic point of view institutions can be defined as 'stable patterns of action that are socially enforced'.[6]

This definition has two components. On the one hand a stable pattern of action has to be in place; in our case we focus on patterns of use. An artefact that is never used by someone for a certain purpose will not have that proper function.[7] But actual patterns of action are not sufficient to establish the normative force of a proper function. Take for instance the fact that we routinely use chairs to stand on. This does not make it into a proper function. Beth Preston calls this kind of secondary use, use according to a system function, '(culturally standardized) *ongoing system functions*' (Preston 2000: 31-33). The point is that this type of use has no normative implications that are special to the use of that artefact. Only general norms, such as the standards of rationality or other (social or moral) norms apply. If we compare this with a regular proper function, the difference is apparent. The normative implications of failed use differ strongly. If my car doesn't start, I have no reason to blame myself, but may blame the manufacturer (or the mechanic who repaired my car recently) (cf. Franssen 2006). Here I assume that the car is used properly, in the sense that it is used according to its proper function and in the proper way. If I fall off my chair, however (for instance because it was a swivel chair) I have no one else to blame but myself. I cannot invoke any further norm.

The normative component is added to this pattern of action with the introduction of 'social enforcement'. This stands for all sorts of social consequences and sanctions that are relevant to the use of artefacts and of the consequences of a particular use. If we look at the examples given above we notice the following. If I use a car properly, i.e. in accordance with its proper function, I am justified on those grounds in my expectation that the car will bring me somewhere. I enter into an institution in which I have certain rights and expectations. And if the car doesn't work, if it malfunctions, then I have several rights. The counterparts of these rights are obligations of others. This can mean that I can hold the manufacturer (or reseller) liable for damage to me or to others. This is the kind of social en-

[6] A more detailed analysis of this institutional view can be found in (Scheele 2005a; 2005b).

[7] This is a rough statement, of course: archaeological artefacts, for example, are *no more* used in their original way. We might want to maintain that these artefacts still have their original proper function. We discover this often by studying the society in which the artefact apparently played a role.

forcement that is connected to institutions and in that sense to proper functions; their (justified) ascription creates obligations. This is not the case with system functions. Rather the reverse: if I intend to use an artefact 'improperly' this can be fine, if it is effective, that is. But if something goes wrong, I cannot transfer any responsibility.[8] Normal standards of rationality will apply here and not living up to those standards will be the user's responsibility.

3. Raz on second-order reasons

Reasons drive our actions; in so far as these actions are intentional. Understanding and evaluating artefact use thus involves understanding the intentions of the user, but also the physical environment in which an artefact is used, and the proper function of the artefact. In the previous section I gave an account of these proper functions that can be summarised as follows. The proper function of an artefact is that *what the artefact is for*. This can be analysed by identifying various conditions for the justified ascription of the proper function. These conditions should involve the physical structure of the artefact, although with some qualifications due to the possibility of malfunction. These conditions also involve various forms of intentionality. The most general way to state the relevant intentions determining proper functions is by saying that they create the institutionalised use of artefacts. The proper function of an artefact is thus partly determined by social institutions within the group of its users. It is possible that an artefact (or an artefact-type) has different proper functions amongst different groups of users.

The question here is how these social norms determining the proper function can help us understand and evaluate the use of artefacts by agents, knowing that we need to combine them somehow with other types of norms, such as rational norms, which are relevant to understanding and evaluating artefact use as well. Thinking of reasons for actions in connection with these norms will help in this matter.

[8] The situation is generally somewhat more complex than this. It might be the case that a kind of alternative use of an artefact should have been possible. The physical structure that enables its proper function can justify the thought that it enables some other system function and the impossibility of using an artefact according to some system function can justify the thought that it cannot be used in its proper way as well (e.g. if I can't use a chair to stand on anymore, I'll probably be unable to sit on it as well). It is not clear beforehand where the responsibility might lie in such cases. Social factors will play a role as well, but not social factors that directly determine the proper function, which I am interested in here.

I propose to view these norms as second-order norms in the sense of Joseph Raz' analysis of norms in action. These norms do not function directly in the deliberation about some action, e.g. by changing the preferences of certain means-ends combinations, but they rather change the decision situation by changing or limiting the allowed options for choice. To see this point we may understand it as follows in a preliminary way. Understanding an action involves understanding the situation an agent is in, or actually, understanding the situation an agent *believes* he or she is in.[9] In a given situation an agent may observe (or think up etc.) a number of alternatives for action. In standard rational reconstructions of actions these alternatives get assigned a preference and an action is said to be rational if the action conforms to the highest preference.[10] However, a second-order reason does not change the preference(s) of opportunities of action, but rather influences the allowable alternatives of choice. It changes the decision situation, because the allowable, as opposed to the preferred options are limited.[11] Raz calls these types of second-order reasons, *exclusionary reasons*: they exclude certain options from the decision matrix: 'An exclusionary reason is a second-order reason to refrain from acting for some reason.' (Raz 1975: 39). As we shall see, the case is slightly different with respect to proper functions. I will extend the analysis to cover these cases, however.

The idea of these second order reasons is as follows. Justification of action forms a central component of the idea of practical rationality. This can be formulated in terms of a 'practical principle: P1. It is always the case that one ought, all things considered, to do whatever one ought to do on the balance of reasons' (Raz 1975: 36). This is one possible formulation of many that have been given in this field of

[9] I use a simple model of action in terms of (rational) belief/desire psychology. Beliefs about the situation (in combination with the desires an agent has) are motivating factors for action, not knowledge or 'the real world' per sé.

[10] I disregard all sorts of details about the preference formation and analysis, but those details are not relevant for my main argument. The formulation used here uses a maximising approach to practical rationality, but it is not difficult to fit it with a satisficing analysis, for instance.

[11] One note. It might be argued that the options that are excluded in this way are simply assigned preference *zero*. This might be the case and also a good way to model it mathematically in a decision-theoretic analysis. However, this does not help in understanding the way agents come to their decisions and what the right reasons and motivations are for this particular preference assignment. Therefore that strategy is not at all useful for us. The distinction between allowable and preferable is a real distinction.

research. For our purposes it is interesting how the author adds exclusionary reasons to this idea. Exclusionary reasons are not part of this 'balance of reasons', but are used in a second principle: 'P2. One ought not to act on the balance of reasons if the reasons tipping the balance are excluded by an undefeated exclusionary reason.' (Raz 1975: 40).[12]

This type of reason may overrule rational actions in certain contexts and be itself of a social (or moral) nature. An example Raz gives is of a soldier who is commanded by his officer to appropriate a van that belongs to a citizen. The authority of the officer is an exclusionary reason that overrides much of the deliberation of 'the balance of reasons' that belongs to a full justification of the action, e.g. that you are not normally supposed to take away someone's property. This authority is a social authority (and a legal authority). This reason is not a 'rational' reason in itself, but it does structure the options of choice of the soldier. It is important to see that an exclusionary reason is not absolutely overriding, but only conditionally overriding (it concerns an '*undefeated* exclusionary reason') (Raz 1975: 38 & 40). The soldier might, for instance, have an even 'higher order' reason *not* to obey the officer. This theory gives us a framework for the evaluation of action in specific social contexts.

This analysis of social norms fits well with the view on institutions briefly described above. Social institutions are the general social structures, which are formed, *inter alia*, by the norms that prevail in society. These norms can be of different kinds, as we saw in the example of the soldier: social, ethical, authoritative etc.

The secondary exclusionary reason need not be a direct order. It can be a standing practice in society and/or be embedded in the legal system. Take the following example in (Dutch) contract law. I can make a deal with someone, which brings a contract into existence. There are different ways in which this contract can be brought into existence or be materialised. I can have a spoken agreement with someone, but I can also write the contract down and both parties can sign it. There are different reasons to choose for one or the other option. If I buy a standard item in a shop it is not necessary and impractical to draw up a complete con-

[12] P1 calls for a universal observance of the balance of reasons, whereas P2 gives a condition under which this should *not* be done. Under that formulation this leads to a contradiction and calls for modification of the first principle: 'P3. It is always the case that one ought, all things considered, to act for an undefeated reason.' (Raz 1975: 40).

tract. For other types of agreements I might want to have a written contract, though, for purposes of administration or for future evidence of the contract, e.g. in the case of problems. These reasons are all first order reasons, i.e. they are part of the balance of reasons in Raz' terminology.

However, take a look at the following example. If I want to buy a house (in the Netherlands) I agree to buy a house and thus make up a contract with the owner. In this case, though, there are strict rules pertaining to the form of the contract that have to be observed. In addition, purchases of this kind and ownership of houses have to be registered in special registers. These (legal) rules exclude other ways of buying and selling houses, even if other ways and forms are possible. In that sense these rules are exclusionary reasons, because they determine or influence the allowed set of options in a given case.

It should be realised that the point is not that it is *impossible* to buy or sell a house in a different way, nor is it, necessarily, a reason that simply changes the balance of reasons, i.e. is part of the calculus in preference formation. A kind of contract that does not conform to the rules given in the law can be valid or can become valid (if it is not explicitly nullified in time). There are several reasons to abide by this rule: you are supposed to observe the law *simpliciter*, but it can also be practically troublesome not to follow this rule.

This is another example of a rule that provides an exclusionary reason, namely by excluding other forms of contracts from the set of allowable actions in buying a house. It does not do this by making it actually impossible, but also not by simply changing the preferences of an action; it comes before these preferences, as it were, and it works differently from means-ends calculus.

An analysis in terms of secondary reasons, and more in particular exclusionary reasons, provides a tool for a more differentiated and thorough understanding of action. It also provides a tool for a more differentiated way of evaluating actions. Following and ignoring exclusionary reasons provides different types of culpability and exculpations. Ignoring the command of an officer has different consequences from not behaving in a rational way (or the most rational way); if only because different 'authorities' will evaluate your case. Although blindly following orders does not exculpate you automatically, it does provide reasons for shifting the responsibility for actions on someone else. These are some of the consequences of Raz' analysis.

4. Extension to *use* of artefacts

Supposing that Raz' analysis sheds light on action in general we can try to extend the analysis to deal with the special case of artefact use. As was said above, much of the relevant norms in this case are provided by the proper functions of artefacts, which are partly socially determined. Do proper functions also provide exclusionary reasons?

I will argue that up to a certain extent they do, but some modifications need to be made. As we shall see the notion of 'exclusionary reasons' should be understood in less strict terms, rather as (second order) *enabling* reasons, which I shall call *institutional reasons*, in the spirit of my view on proper functions explained briefly above.

The ascription of proper functions in terms of institutions shows how social norms are relevant to artefact use. On the one hand an artefact has many ways that it can be used for rationally: these are its system functions. Artefacts are indeed generally used for many different things: this may be one-off use (quickly using your mobile phone as a paperweight to prevent papers being blown away); this may be an accepted (stable) pattern of use (using a chair as a step ladder), but there is no special normative force connected with these uses. The normative force comes into play when we add the institutionalisation of this use, whatever the source of the institution may be. In normal cases, i.e. in cases where the reason is not defeated, this institution forms a reason to use the artefact in that particular (proper) way. If the artefact malfunctions (and if this is known), we can say that the reason is defeated.

The way in which a second order reason, in this case through the institutionalisation of use, works here is by providing a natural or standard way for doing some job. If we want to have a hole for a screw in a wall, we shall immediately think of a power drill, because that is what those things are for. The proper function of a power drill thus helps structure the decision space in this case. In addition this institutional reason also helps judging failed use. If I use the drill in a correct way but the drill fails to do the job there is reason, on the grounds of the proper function of the drill, to blame the manufacturer, designer or reseller. The proper function of an artefact makes that I can have justified expectations about the operation of the object. These expectations are based on certain social norms that are associated with the proper function of the object. These social norms, part of the institution, form second order reasons for performing actions.

So far, the analysis of proper functions provides norms very similar to those that provide exclusionary reasons. However, an important difference with Raz' analysis that is connected to viewing proper functions as institutional reasons concerns the fact that proper functions in fact are not simply *excluding* reasons. The social institution does not really exclude other uses, but it rather gives reasons for performing an action one way, rather than another. In contrast to exclusionary reasons, there is no real (social) problem with use according to system functions (barring irrational use). Especially creative alternative uses are often judged positively. Strictly seen, exclusionary reasons only allow for such normative freedom if they are *defeated* (because of some other, overriding norm), but for many cases of alternative use this needs not be the case. What the proper function *does*, among other things, is structure the *consequences* of use, most notably in cases of failure. The consequences of failure in cases of proper use differ from the consequences in cases of improper or alternative use.

An important additional difference between exclusionary and institutional reasons concerns the possibility of holding some party responsible or liable in the case of alternative use (which is especially important if an accident happens, of course). If an artefact is used according to its proper function, there is reason to suppose that the producer of the artefact is responsible for negative (side) effects or accidents occurring when using the artefact.[13] If I misuse an artefact or use it 'improperly' there is every reason to blame me for failed use and/or accidents. This is the kind of difference that is relevant for the way proper functions have normative force, as institutional reasons. These different results are also, for instance, institutionalised in the form of laws in a country or warranty certificates that come with products.

Our analysis of reasons should deal with cases of failure. Proper functions structure the decision situation. They don't do this directly, but by changing the risks (of failure of use). I would say that these are second order reasons as well, but differently from exclusionary reasons. For that reason I introduced the term '*in-*

[13] This should be qualified. If I have an accident with a car when driving it, it very much depends on the circumstances whether I can blame the manufacturer (if all else is in working order). The same is true for using a gun. Cf. the NRA slogan "Guns don't kill, people do" in its campaign to keep gun possession legal and to prevent gun-producers to be held liable for killings and accidents with guns. Opinions differ about the consequences of accepting this statement, of course.

stitutional reasons' for the normativity that is introduced by proper functions. By using an artefact you become part of an institution and you also *reaffirm* the institution. The opposite can also be the case. By using an artefact in an alternative way you place yourself outside the relevant institution and it can even be a means to undermine the institution and be a force in a process of institutional change.

Different types of second order reasons now may be seen to interact in different ways, sometimes mutually enhancing each other's norms, sometimes conflicting. Take for instance the second order reason to use an object for a certain purpose, e.g. a gun is properly used to shoot with and in that sense you make yourself part of a certain institution with its institutional norms. On the other hand it is deemed immoral to use guns to shoot people in most contexts. This can be interpreted as a social exclusionary reason *not* to use the gun in that way. This causes a conflict of norms. Norms of rationality may come into play again, for it might be or might not be rational to conform to some social institution. And this norm of rationality may again be thought to be overruled by some higher social norm. This is not the place to investigate all these levels of normativity, because they no longer concern artefact use in a strict sense. The example only serves to indicate that the concepts of exclusionary reasons and institutional reasons as secondary reasons serves an important analytical purpose in the investigation of artefact use.

5. Conclusion

In this article I have argued the following. The social aspects of proper functions can be understood in terms of institutions. Use of artefacts according to the proper function is termed proper use. Proper use, in turn, is embedded in the normative structure of social institutions, i.e. those that help determine the proper functions of artefacts. These social normative structures are complementary to traditional standards of practical rationality and are a kind of *second-order* reasons, similar to so-called exclusionary reasons. Proper functions provide *institutional reasons*, which are in some respects different to these exclusionary reasons. The most notable difference concerns the fact that proper functions not so much exclude other types of use, but rather place that use (and the user) in different social structures with different rights and obligations. An institutional reason not only gives a reason for action, it also provides reasons for evaluating actions according to such reasons positively.

The upshot of this analysis is that it gives us an additional tool to understand and evaluate the use of artefacts. This tool provides for a more differentiated and thorough *understanding* of artefact use. It also provides a tool for a more differentiated way of *evaluating* actions. We can use artefacts according to their proper functions, and we usually do, but we need not do this. Proper functions provide second order reasons for a certain kind of behaviour, but they do not force this behaviour.

Institutional reasons show how there are differences between proper use and other kinds of use. These differences become most clear when some action goes wrong. As is the case with exclusionary reasons, acting in accordance with a proper function, and thus in accordance with an institutional reason, can work exculpatory when something goes wrong. This means that we can or should evaluate such actions differently from use that is not done for an institutional reason. The analysis is not just relevant for the evaluation of use, it is also relevant for understanding use: i.e. use is not just done "on the balance of reasons", but rather because some artefact simply is supposed to be used in such and such a way.[14]

Bibliography

Buller, D. J., Ed. 1999. *Function, Selection, and Design*. SUNY Press: Albany (NY).

Franssen, M. 2006. The normativity of artefacts. *Studies in the History and Philosophy of Science* 37: forthcoming.

Houkes, W. and P. E. Vermaas. 2004. Actions versus Functions: a Plea for an Alternative Metaphysics of Artefacts *Monist* 87 (1): 52-71.

Millikan, R. G. 1999. Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function. *Journal of Philosophy* 96: 191-206.

Neander, K. 1991. Functions as Selected Effects: the Conceptual Analyst's Defense. *Philosophy of*

Science 58: 168-184.

[14] This article is based on part of my dissertation (Scheele 2005a) and was presented at the SPT Conference: Technology and Designing, July 20-22, 2005, Delft University of Technology. I would like to thank those present for their useful comments and criticism. In particular I thank Anthonie Meijers and Wybo Houkes for their comments at the conference and on earlier versions of the article.

Perlman, M. 2004. The Modern Philosophical Resurrection of Teleology. *Monist* 87 (1): 3-51.

Preston, B. 1998. Why is a Wing Like a Spoon? A Pluralist Theory of Function. *Journal of Philosophy* 95: 215-254.

Preston, B. 2000. The Functions of Things, a Philosophical Perspective on Material Culture. In *Matter, Materiality and Modern Culture*, edited by P. M. Graves-Brown. London, Routledge: 22-49.

Raz, J. 1975. *Practical Reason and Norms*. Hutchinson & Co: London.

Scheele, M. 2005a. *The Proper Use of Artefacts. A Philosophical Theory of the Social Constitution of Artefact Functions*. Marcel Scheele: Leiden (isbn: 90-9019521-1).

Scheele, M. 2005b. Social Facts from an Analytical Perspective. The Example of Institutions as a Unifying Notion in the Social Sciences. *Graduate Journal for the Social Sciences*: pp. 101-127 (www.gjss.org).

Scheele, M. 2006. Function and Use of Technical artefacts; the Social Conditions of Function Ascription *Studies in the History and Philosophy of Science* 37: forthcoming.

Function and Probability: The Making of Artefacts

Françoise Longy
Institut d'Histoire et de Philosophie
des Sciences et des Techniques, Paris

Abstract: The existence of dysfunctions precludes the possibility of identifying the function to do F with the capacity to do F. Nevertheless, we continuously infer capacities from functions. For this and other reasons stated in the first part of this article, I propose a new theory of functions (of the etiological sort), applying to organisms as well as to artefacts, in which to have some determinate probability P to do F (i.e. a probabilistic capacity to do F) is a necessary condition for having the function to do F. The main objective of this paper is to justify the legitimacy of this condition when considering artefacts. I begin by distinguishing "perspectival probabilities", which reflect a pragmatic interest or an arbitrary state of knowledge, from "objective probabilities", which depend on some objective feature of the envisaged items. I show that objective probabilities are not necessarily based on physical constitution. I then explain why we should distinguish between considering an object as a physical body and considering it as an artefact, and why the probability of dysfunction to be taken into account is one relative to the object as member of an artefact category. After clarifying how an artefact category can be defined if it is not defined in physical terms, I establish the objectivity of the probability of dysfunction under consideration by showing how it is causally determined by objective factors regulating the production of items of a definite artefact type. I focus on the case of industrially produced artefacts where the objective factors determining the probability of dysfunction can be best seen.

Keywords: artefact, probability, etiological theory of function, biological function, artefact function

Function and capacity

One usually associates function with capacity. Coffee machines usually have the capacity to make coffee and hearts, which have the function of pumping blood, are usually able to pump blood. This relationship is of practical importance : often, it is by learning the function of an object that one learns what to do with it and what to expect from it.

One of the two major contemporary theories of function, that goes back to an article of Cummins in 1975, relies on this close relationship : it identifies the function to do F with the capacity to do F or, to use a more technical term, with the disposition to do F. But by doing so, it offers no account of the normative aspect of functional discourse. An entity that has a function is supposed to do something under particular circumstances, but it may not necessarily be able to do it as dysfunctioning items (a non-working coffee machine or a diseased heart) show. We will set aside here the question of whether there are two types of functions, the first one having normative import and allowing us to speak of dysfunctions while the second does not. We will concern ourselves only with the first type of functions.

The other major theory, the so-called etiological theory, whose basic tenets go back to an article of Larry Wright in 1973, offers a straightforward account of dysfunction, which is a reason for its wide acceptance by philosophers. According to this theory, functions indicate a particular sort of history, and that explains their normative import as well as their etiological sense. Actually, functions often serve to explain why something exists or is so : the function to attract peahens is usually supposed to explain why peacocks have such a big and vivid tail. For the etiologists, in fact, a function is an effect that explains a “being there.” For instance, the current actual hearts have the function of pumping blood, because previous hearts have been selected for having had the effect of pumping blood. In view of their dependence on past facts, functions can be dissociated from present capacities. The possibility of dysfunctions rests on this temporal discrepancy : something can have a function because of its history, even if it fails at present to have the corresponding capacity.

Some etiologists have remarked, furthermore, that the relationship between function and present capacity cannot simply be of the form that to have a function implies a high probability of having the corresponding capacity. Often the two go together, but this is not necessarily the case. To show it, Millikan contemplated the case of the spermatozoids. They have the function of fertilising ovules, but that does not imply that they have a high probability to do so. Neander has put forward, in the same intent, the case of a pandemic disease. If a pandemic disease would make 75% of the population blind, the function of the eyes would still be to allow vision. This discrepancy between function and capacity has been a further argument to exclude present capacities from the notion of functions, and so far that is what etiologists have done.

The drawbacks of the current dualist theory of function

Up to quite recently, etiologists have been mostly interested by biological functions. So it was natural that they were ready to define functions by referring to natural selection. Some extended this definition to the functions of artefacts, by admitting not only natural selection but also cultural selection. But this can be only a partial answer to the question since most artefacts are attributed a function when they are created, i.e. before any cultural selection could act on them. The etiologists who tackled this point answered that, in this case, the function names what the person(s) responsible for creating or producing the artefact thought it was for.

But this thesis, here called the intentionalist theory of function, has a severe drawback. Such "intention-based" functions are, in a specific sense, subjective: they will vary according to the intentions attributed to their designers, all other things being equal. For example, let us consider a component that has been made and put in some specific place by Boris, it will have the function to cool the liquid passing through it if it is what Boris thought it was for, or it will have the function to reflect the incoming light if that was Boris' idea. No objectively ascertainable factor need vary from one situation to the other for the function to vary : the change of the intentional content in Boris' mind is sufficient.

Such a direct dependence on intentions must be clearly distinguished from the dependence on intentions through socio-cultural pressure which may appear in a cultural selection theory of function. It is a general truth that the use of an artefact and its diffusion depend essentially on what people think about it : what it may do, how it should be used, what it could be useful for, etc. But such intentions will be taken in account in a cultural selection theory of function only when they manifest themselves and thus contribute to the general social-cultural context determining the “evolutionary life” of the artefact. Now, a cultural context is as objective as a natural one. One looks at what people do and did : how they use(d) the artefacts, the preferences they show(ed) in buying them, etc. This is something that can be studied by empirical means, the ones history and sociology currently draw on with no need for introspection. Therefore, the intentionalist theory of function must be clearly distinguished from the thesis that artefact functions depend on intentions through socio-cultural pressure and selection. Only the first one endows functions with a particular subjective nature.

I have argued elsewhere that the intentionalist theory has to be rejected (1) because no artefact function manifests the subjectivity this theory implies and (2) because it would imply a highly problematic ontological heterogeneity : two very different sorts of properties - subjective ones and objective historical ones - would be indistinguishably mixed. The conclusion I have drawn from these considerations is that the classical dual etiological theory of function – selectionist for biological items and eventually a large part of the artefacts and intentionalist for the remaining part of the artefacts – is mistaken and that one must look for a more general and more abstract characterisation of etiological functions.

Theory of function : A new perspective

The challenge, then, has been to see whether and how one could accommodate the two subsequent facts while maintaining an etiological perspective :

1.There are artefact functions which are not due to a selection mechanism
2.All artefact functions are objective.

In other words, which objective property could explain why artefacts are there, whether they have just been created or have long been maintained by cultural selection?

A designer conceives an artefact, i.e. a type of artefacts, for a determinate function : it should have the capacity to do F in a determined type of circumstances. So, the objective property we are searching, let us call it O, should be related to this specific capacity. But to go along this line implies overcoming two serious difficulties.

First, how could there be any objective property of the Xs, prior to the existence of any X, that could explain the existence of Xs? What is required is an O such that “X is there because of O”. It has usually been thought, that O should be a past event or a past state of affair. It is not necessary. O could be a timeless property – as is, for example, the relation between the type X, the capacity to do F and a series of circumstances C – and play a part as a reason. For instance, someone finding out about O decided to make Xs because of O. We reserve for another article a proper defence of this way of understanding the etiological condition attached to functions.

Second, how to justify the probability that then has to be introduced in the definition of function ? The necessity to introduce probability comes to light when one tries to answer the following question : What objective link can exist between the Xs and the contemporaneous capacity to do F when the Xs have function F ? As we have seen before, an item can have a function and lack the corresponding capacity. So, the function to do F can *at most* imply some probability to do F. Consequently, if a definite relation exists between the function F and the capacity to do F, it can be only probabilistic.

The aim of this paper is to show (a) that there is in fact for each function F a specific probability of having the capacity to do F (the same generic high probability for all cases will not do, as spermatozoids show) and (b) that this introduction of probability is no smart trick but rests on solid grounds. It is not good enough to have the valid negative reason that functions cannot be equated in a straightforward manner with their corresponding capacities. Introducing probability to loosen a tie that would otherwise be too tight is *ad hoc*. Other reasons must justify it positively.The line of thought summarized above led me to the following characterisation of function : "X has function F" means :
As an item of Type ***X***, X has a(n) (objective) property O such that :

1)O implies that X has probability P to do F in circumstances C

2)The present X or Xs are there because of O

Let us outline two aspects of this characterization to show how it can cover artefact functions as well as biological ones. First, O can correspond to properties of very different sorts. In the case of a biological function, O will be roughly: belonging to a species some of whose previous members have been selected because they have had an X that did F. But for the first generation of an artefact, O may be something like : an object constructed in such a manner (in order to fall within some margin of error this series of specifications) will have, because of the laws of physics, chemistry, etc., probability P to do F.
Secondly, the mechanism or the process which explains causally the existence of the Xs does not enter into this characterization (in that it differs from most

etiological definitions of function). It can rely on selective forces or on intentional contents and actions insofar as the two conditions above are satisfied.

Finally, to avoid any confusion, let us make precise that the characterization of function we propose here is quite different from the one Bigelow and Pargetter proposed in 1987, although they also introduced probability. For them, a function is a capacity enhancing the fitness (i.e. the chance to survive) of the entity possessing it. There are two major differences between their proposal and ours. First, according to them, functions inform us about the future not about the past. Second, the probability they introduce concerns not the possession of the capacity itself but the survival value resulting from possessing the capacity.

After this general presentation, let us turn our attention to the more specific questions that are the target of this paper : "What does the probability appearing in the first condition consist in?" and "Does this probability have real grounds that justify it positively?" To answer them we need first to make clear, in general, when probability will correspond to something real and substantial and when it will not.

When do probabilities reflect arbitrary conditions ?

If you say of some 39 years old Parisian woman that her probability to be pregnant today is P using the information that women between 35 and 45 who reside in Paris have probability P to be pregnant in this period of the year, you point to an objective feature relative to the category (the ratio of pregnant women among them) but not relative to the woman. Relative to her it just reflects your level of information and the perspective which you adopt in looking at her. It is because you consider her as a member of this particular group that you attribute to her this probability to be pregnant today. With new information, for example, that she suffers from some gynaecological problem, or that the ratio of pregnant Parisians between 38 and 40 is P', the probability would change. New information means a new reference class, and this implies generally a new probability value. Better knowledge may even allow us to dispense with probabilities. It would be the case, if you came to know through a sonogram that a fecundated ovule has nested in her uterus. We will call such probabilities perspectival : they depend on an arbitrary perspective, a perspective determined simply by a pragmatic interest or a particular state of knowledge. The criterion for recognizing these probabilities is that they would change and may even disappear if the arbitrary limits imposed by our current knowledge change. On the contrary, we will call “objective” the probabilities which can be shown to be independent of an arbitrary perspective due to a provisional state of knowledge or to some pragmatic interest.

A simple case of objective probability is the one associated with non-linear dynamical systems. These systems are highly sensitive to initial conditions : any small variation in the initial conditions can give rise to enormous differences, the so-called "Butterfly effect". So, whatever the precision you may obtain in fixing a range of initial conditions, completely different outcomes will remain possible after a long enough dynamical evolution. No increase in knowledge would make it possible to get rid of probability in predicting the outcome after a long enough dynamical evolution. The probability is tied to a physical property of these dynamical systems : their high sensitivity to initial conditions. But objective probability can also be grounded in something not purely physical. To demonstrate this, let us take the example of throwing a die.

Here there are two phenomena susceptible to giving rise to an irreducible probability. The first one is the physical phenomena we have just seen. The high sensitivity to initial conditions of rolling dice will imply the equiprobability of halting on any face after a determinate number of rotations, let us say after 10 rotation. The second one is the social phenomena of using dice in chance games. It is clear that on less demanding conditions the outcome of a rolling die may be quite predictable. For example, if someone puts the die in her hand with the 4 on top and makes the die slide to the table without rolling, the probability that the die will halt on 4 will be 1. Maybe the probability to halt on 4 will be 1 also, if one makes the die roll only once from a determinate initial position in the hand. It is because one easily sees that such ways of "throwing" dice makes the outcome quite predictable that no player is allowed to make dice roll only once.

However, there may be ways of throwing dice that are allowed, even if their result (the halting face) could be predicted in a categorical way were one to know the value of some parameters (initial position, force of throw, etc.) with a determinate precision. For example, let us suppose our physical theory makes it possible to predict on which face a die would halt if it rolls only three times on a plane from a determinate range of initial positions. It may well have no importance for the fairness of a dice game played by human beings, even if the players are physicists in possession of a table giving them the different outcomes for a series of ranges of initial conditions. Why? The limited capacity of bodily control humans have may imply that they necessarily pick at random in a large range of initial conditions when throwing, and this, in turn, may imply the same chance for every face to come out after only three rotations of the die. If it is so, the probability of 1/6 to halt on 4 can also be objective when the die has rolled only between three and nine times and was thrown by a human being : it is an objective probability related to the real use of dice, their use in chance games. In fact, the dice, the table and the person who throws may be seen as a new sort of dynamical system. The point then is that the initial conditions take in account the fact that it is a very complex machine, a human, who is throwing : the human hand or arm cannot be isolated and treated as a simple mechanical device whose physical parameters could be set up at will.

In our perspective which is to associate probabilities to artefact functions the more interesting case of objective probability is the last one where social factors (dice as used in human chance games) are taken in account. Why do social factors matter when considering objective properties of artefacts ? We will once again use the dice as a paradigmatic example to answer this question.

Artefacts and causal explanations

The probability of 1/6 to halt on 4 after only three rotations is not a physical property of the rolling dice as is the probability of 1/6 to halt on 4 after ten rotations. The latter one is a result of physical theory when considering dice purely as physical objects : it is the probability obtained by considering situations where one could imagine to fix as precisely as wanted the pertinent physical characteristics of a system comprising a die, i.e. a well equilibrated cubic object, a horizontal plane and a simple mechanistic throwing device. This probability tells of the sensitivity to initial conditions well equilibrated cubic objects have when rolling. Conversely, the first probability tells something about dice inasmuch as they are part of a specific social situation : chance games played by human beings.

What needs to be highlighted is the following : this type of situation (human dice-throwing in the context of chance games) is not an arbitrarily defined category, it has a substantial reality in our world, a world which is not simply a physical world of material bodies, but is also a world of living beings with social activities. A lot of facts concerning present and future dice depend on the fact that dice throwing is a human game. For objects considered as physical bodies – like dice as well equilibrated cubic objects - all causal explanations are the exclusive concern of physics, but for objects considered as artefacts, many causal explanations hinge on socio-cultural factors. The size of dice depends essentially on the size of human hands. The precision with which they are made depends also on their use in games : they should be sufficiently well equilibrated and of a sufficiently regular cubic form so as to raise no worry about the equiprobability of the different faces to turn up. To sum up, the production of dice is not guided by the concerns of physicists interested in the behaviour of well-equilibrated cubic objects, but by the concerns of players who throw dice on common dining tables at home or on the velvet of gaming tables in clubs.

What do we need to know about dice to make sensible predictions about what may happen to dice ? Not so much what characterizes them as physical bodies as what is their typical use, their function. It is their use that will explain why some are found in children's pockets or why some ended up in some dump or other when their colour faded and they looked old or dirty. Mental experiments show this still more vividly.

Let us suppose, just for a second, that a die rolling on a horizontal plane is a linear system and that our present physical means allow us to predict categorically most of the time on which face a rolling die would stop after 10 or more rotations. Would that change anything? Nothing much if dice throwing remained an activity performed as it is performed today with a limited control of the players on initial conditions, and if this limited control implied an equiprobability to halt on any one of the six faces after three rotations. The physical theory concerning the dynamics of well-equilibrated rolling cubic objects would be completely different but the story of dice could be the same. Conversely, what would induce changes would be the creation of devices, some bionic arm for example, allowing a better control over initial conditions and making it possible that a well equipped and trained gambler could quite often make the die halt on the face she wanted. Most certainly then, the rules for throwing dice in gambling houses would be changed or new sorts of dice would be made, for example dice with an internal rotating sphere introducing a higher sensitivity to initial conditions.

"Dice" names a functional category, not a physical one, and that is no minor detail. A physical category is a category defined by a series of physical properties like, for example, being a well equilibrated cubic object whose edges are between 0,3 cm and 20 cm, while a functional category is defined by a specific use. As we have seen above, one has to take into consideration this specific use if one wants to explain and predict many properties of present and future dice. The function of being a die implies of course the possession of determinate physical properties, but the hierarchy is clear : function comes first. It shows still more vividly with complex artefacts like cars or engines. It is not by extracting the common physical properties of present cars, that one will obtain a good definition for car, a definition that will be able to encompass future cars; this can be accomplished only by considering what they are made for. For instance, knowing that cars are for personal transportation (between 1 to 10 people), one can deduce some property future cars are almost certain to have, for instance seats with sufficient front room for human legs. What is uncertain, on the contrary, is the presence of wheels, even if it is presently a feature all cars possess and have possessed : maybe future cars will use air cushion or quick caterpillar tracks.

Considering artefacts as physical entities or as functional ones

We don't attribute to dice some mysterious properties by saying that the probability a die has to stop on one face after rolling only three times may be different whether we consider it as a physical object or as an artefact. As we have seen, considering a die as a physical object or as an artefact means simply that we are envisaging different sorts of situation. Problems arise only if one fails to notice the ambiguity that phrases of the form "the probability of X to do F in circumstances C" may sometimes have.

Supposing that probabilities correspond to frequencies in some reference class (or population), it is sheer triteness that different reference classes will generally mean different probabilities. But sometimes the reference class is left implicit while different ones could be meant. Speaking just of the probability of X to do F may be ambiguous *when* X *is an artefact* because X can be envisaged, as we have seen, either as a member of a physical category or as an artefact, i.e. as a member of a functional category.

Let us consider simpler examples than dice: artefacts whose capacities imply no probability, "surefire" capacities as Mackie called them (conversely, to be a fair die implies the possession of a probabilistic capacity). A surefire capacity to do F in circumstances C will manifest itself by producing outcome F every time circumstances C are present. For instance, to be water-soluble implies to dissolve whenever put in water in a definite set of circumstances (being on earth, ...). No exception is allowed. If the expected outcome does not show, it proves that the capacity was in fact missing. Electric switches, bulbs or hairdryers are endowed with such capacities : they are supposed to turn the current on and off, to produce light, to blow hot hair whenever a definite set of circumstances is present.

That is no doubt a simplification. Often, the capacity an artefact is supposed to have is related to graduated effects, and that means that considerations of level and border line effects step in. A hairdryer that blows very little air will be judged not to have the capacity hairdryers should have and will be counted as a dysfunctioning hairdryer. Some, the border line cases, will appear as not working perfectly, blowing not exactly enough air or a bit too much. The notion of well-functioning often goes with the existence of standards : the capacity should result in effects that exceed some limit or are in between definite values. But we can, by supposing a standard precise enough and very rare border line cases, ignore here these complications so as to focus on our major question : what is the probability to do F about when we consider an artefact relatively to the function of doing F.

Let us consider the case of the bulb. The capacity at issue is the capacity to produce light when connected in the right way to the right electric settings with the right amount of current passing through. Which probability should enter in our characterisation of the function of the bulb ? Not the probability the bulb has when envisaged as a physical object. As a physical object, that is as an object a physicist will analyze by looking at its physical structure and characteristics, no probability will in general be implied : whether or not it has the wanted capacity will be a straightforward matter. There may be some irreducible border line cases, for example, when the filament is weak at some point in such a way that it is indeterminate whether it will break down or not when heated by the current passing through. But in the general case, it will be a quite definite matter whether it will produce light or not in the right circumstances, and our present physical means of analysis are already sufficient for giving in most cases a straightforward yes or no answer with a minimum risk of errors.

If what one was considering was a physical category - all objects whose physical characteristics are very close to the ones of this particular X - then it would generally be a straightforward matter, with no need to appeal to probability, whether the Xs of this physical type would have or not the capacity to produce light when placed in the right conditions. But this is not what we are interested in when considering artefacts. What interests us is whether the object produced or bought as a determinate artefact will have the desired capacity. The question then is not whether a determinate physical structure, which is instantiated by this particular bulb, has a determinate capacity or not but whether or not an item belonging to the category bulb has a physical structure allowing it to have the desired capacity.

Functions do not necessarily imply multirealization, as it is sometimes supposed, but they go happily with it. Being a bulb, a can opener or an engine supposes some specific relation to a particular capacity (a relation more complicated than simply possessing the capacity, as we already know) but it does not imply a determinate physical structure. Several physical structures can be found in the same functional category. (I will consider below the intriguing question of how functional categories can be defined.) Thus, two quite different things can be hidden in the too general phrase "the probability X has to do F in circumstances C" : on the one hand, the probability the physical structure X has to do F and, on the other hand, the probability the functional item X has to possess a physical structure doing F. But in order to perceive the ambiguity and to be willing to eliminate it, one has to be convinced first that it makes sense to distinguish the functional object from the physical one, and that this distinction is required for explaining artefacts – what may happen to them, how they may change, etc.

The difference we are stressing here is quite similar to the one that can be found when speaking of organisms, where the same sort of ambiguity can be encountered as well. "What is the probability that a particular baby becomes an overweight child if she has this diet and performs these physical activities ?" may mean "what is the probability that she becomes overweight since she has this particular genetic make-up ?" or "what is the probability that she possesses a genetic make-up driving her to get overweight since she belongs to this particular population or has had these ancestors ?". In biology, such ambiguity appears to be quite common and goes together with the existence of two different sorts of causal explanations for the same phenomenon, one pointing to physiology or development and the other to heredity or evolution. It is with the intent to account for such a duality that Mayr introduced the distinction between "proximate causes" and "ultimate causes". The very nature of evolutionist explanations as well as the relation these entertain with physiological or developmental ones are still discussed issues in the philosophy of biology. Without tackling any of these questions, one can just notice the legitimacy of another level of causal explanations for organisms than the one of proximate physical causes. What we defend here is simply that a similar duality has to be admitted in the case of artefacts too.

Design, industrial production and the probability of being a well-functioning item

How are functional categories defined if they are not defined in physical terms ? By historical factors, like species are defined. In general, items of the same artefact type have been produced in the same industry or in similar ones, following identical procedures or following procedures that have been seen or demonstrated to give rise to identical or similar outcomes, they have been submitted to identical or similar controls relative to the same properties, they have been distributed, offered for sale, advertised as objects of the same functional type. Furthermore, the manufacturing processes will often result from common engineering and design processes or from ones that are largely related to one another. This is what links together items of the same artefact kind, when one looks at them from the production side. On the other side, the consumers' one, the functional identity is currently perceived and maintained : objects bought or transmitted as items of the same artefact kind will be used in the same ways and will be expected to behave identically in circumstances related to their typical use. As soon as there will be different trademarks, the buyers will act as a selective force making the trademark with more dysfunctioning items disappear or cost less or improve their products. The members of a same artefact kind are not linked together as strongly as the members of a same species are - heredity through transmission of genetic material – but their linkage is sufficient to determine a real kind of a historical nature.

The probability artefact X has to be a well-functioning item - the probability that X has to possess a physical structure doing F when doing F is the function of its artefact type - can be evaluated by statistical means : the proportion of well-functioning or dysfunctioning items in representative samples of the relevant population**.** But that it can be so evaluated does not say anything about the nature of this probability. It does not tell whether it reflects only epistemic or pragmatic factors, or has a causal ground. In other words, it does not tell whether the probability is perspectival or objective. We will try to show that the probability is objective when it is calculated relative to a real-kind population and fulfils determinate conditions.

The probability to possess the required capacity is an explicit pivotal element in the industrial production of artefacts. It already plays a role at the stage of engineering and design. Engineers envisage artefacts in such probabilistic terms when they work out their specifications and how to realize them : which materials to use, what should the production line be, which controls to perform at which stage, etc. The choices engineers or managers have to make usually take into account how doing one thing or another will increase or decrease the probability of dysfunctioning items. For instance, in order to minimize production costs while remaining under the threshold of 0.5% of dysfunctioning items, will it be better to buy cheap materials and install at some stage of the production line a device eliminating 98% of the defective elements or to buy expensive high-quality materials ?

Usually the value of the threshold that the proportion of dysfunctioning items should not exceed is explicit. What sets this value ? Mostly the competition on the market and the consequences dysfunctioning items may have. So, through the retroaction of the market on the production, this threshold - and hence the probability to be a well-functioning item if manufactured in this country or under this trademark - will depend on social factors like which reliability-price ratio will help to ensure good sales.

To sum up, the probability an artefact item has to be a well-functioning one depends on the conditions and processes of its manufacture, and these, in turn, depend for their maintenance or improvement on a great many factors : technical discoveries, costs of possible improvements, expectations about quality, expectation about costs, level of commercial competition, etc. In other words, when X is considered as a member of the population of artefacts a specific factory produced within a period of stable manufacturing conditions, its probability to be a well-functioning item reflects some objective features of the complex causal process responsible for producing the entire population, and not some arbitrary perspective under which X would have been considered.

There are different reference classes or populations that satisfy the requirement of resulting from a common origin or from causally interdependent origins in such a way that they match a causal process capable of explaining many features of their members. To try to precise more formally how to characterize these populations will raise very difficult questions like the one of defining real kinds. An intuitive grasp, that can be tested on examples, is sufficient here. Hairdryers coming out of one factory, hairdryers of a particular trademark (the same manufacturing standards will be imposed to all production units), or hairdryers of well-established occidental trade-marks distributed in occidental countries are all examples of such real-kind populations. Conversely, the population of yellow hairdryers produced in May 1999 in Singapore has no reason to match a specific causal process having an explanatory power relative to this specific population (of course to have been painted in yellow will explain their being yellow, but this causal relation is almost a tautology, it cannot be said to have any significant explanatory power).

Real-kind populations like the ones we considered above can be part of one another or can even sometimes overlap. Is that not a problem for the position we defend here ? It is no more a problem in this case that it is a problem in biology to explain some phenomena considering levels under or above the species level. For instance, it may be pertinent to explain the proportion of sickle cells anaemia in some part of the world, and hence a probability to have some specific gene, to a specific subgroup of the human species that has lived in relative isolation in sub-Saharan Africa.

The expected proportion of dysfunctioning hairdryers of well-established occidental trade-marks distributed in occidental countries within the same range of price, let us say $P^{OC,}$ results from what are the expected proportions of dysfunctioning hairdryers in each factory concerned, let us say P^1, P^2, etc. The different P^ks will normally be very close to one another. The P^ks corresponding to factories of the same trademark will be more or less identical because of the standards set by the trademark, the P^ks of different trademarks will be very close one another because of the forces exerted by the market. So, in the end P^{OC} will be very close to the value of each $P^{k.}$ If P^{OC} would be obtained as a simple means between the different P^ks and the P^ks were depending on unrelated factors, P^{OC} would reflect something arbitrary. But the P^ks depend on the market, as we have seen, and that is what P^{OC} reflects : the forces the market exerts on hairdryers' quality in a society having such an economy and such technological resources.

Now, we are in position to offer a general answer to the question raised earlier about whether the probability X has, as an artefact, to be a-well functioning item is objective or perspectival. It will be objective if the artefact population relative to which the probability is evaluated is a real-kind and if the expected ratio of well-functioning items in this population is a consequence of what grounds it as a real-kind, otherwise it will be perspectival. So, for example, whoever supposes that the population of hairdryers considered for evaluating P^{OC} is defined only by a pragmatic interest (like wanting to know which chances there are to buy a well-functioning hairdryer in this price range) should conclude that P^{OC} is perspectival. But, anyone who accepts that such a population is a real-kind and that P^{OC} results from factors determining it substantially should conversely reach the conclusion that P^{OC} is objective.

To conclude, let us sum up our principal result : there is a specific probability or a probability bracket that can be attributed to an artefact item to have the capacity for which it is made, and this can be explained by and grounded on objective factors. By that, we have tried to show that to link a function to the probability of having a corresponding capacity was, in the case of artefacts, not only possible, but also much more than just a technicality since this probability was rooted in the causal processes underlying artefact categories. An ulterior justification of our characterisation, which will be left for further investigations, will be to show why functions, so understood, are at the same time epistemically and causally important : (1) why we find it useful in so many cases (for organisms, for artefacts, etc.) to refer to such functional categories and (2) why this is a way to carve the world at some of its joints so as to obtain valid causal explanations.

Bibliography

Bigelow, John and Pargetter, Robert. 1987. Functions. *The Journal of Philosophy* 84: 181-196. Cummins, Robert. 1975. Functional Analysis. *The Journal of Philosophy* 72: 741-765.

Diaconis, Persi and al. 2004. Dynamical bias in the coin toss, http://www-stat.stanford.edu/~cgates/PERSI/papers/headswithJ.pdf

Houkes, Wybo & Vermaas Pieter. E. (eds.). 2006.*Artefacts in Philosophy* (forthcoming). Peter Kroes, Coherence of structural and functional descriptions of technical artefacts, *Studies In History and Philosophy of Science* Part A, Volume 37, Issue 1, The dual nature of technical artefacts, March 2006: 137-151, http://dx.doi.org/10.1016/j.shpsa.2005.12.015

Longy, Françoise. 2006a. A Case for a unified realist theory of functions, in Houkes & Vermaas P. (eds) 2006.

–––. 2006b. Unité des Fonctions et Décomposition Fonctionnelle, in *Le tout et les parties* edited by Jean Gayon and Thierry Martin, Paris : CNRS éditions (forthcoming).

Lorne, M-C. 2004. Explications fonctionnelles et normativité, PhD Thesis, Paris : EHESS.

McLaughlin, P. 2001. *What Functions Explain*, Cambridge : Cambridge University Press.

Millikan, R. 1993. White Queen Psychology and Other Essays for Alice, Cambridge : The MIT Press.

———. 2000.*On clear and confused ideas*, Cambridge : The Cambridge University Press.

Mayr, Ernst. 1961. Cause and Effect in Biology, *Science* 134, 1501-1506
Neander, Karen. 1991. Function as Selected Effects: The Conceptual Analyst's Defense, *Philosophy of Science* 58: 168-184.

Wright, Larry. 1973.Functions. *Philosophical Review* 82: 139-168.

The (Alleged) Inherent Normativity of Technological Explanations

Jeroen De Ridder
Delft University
Department of Philosophy
The Netherlands

Abstract: Technical artifacts have the capacity to fulfill their function in virtue of their physicochemical make-up. An explanation that purports to explicate this relation between artifact function and structure can be called a technological explanation. It might be argued, and Peter Kroes has in fact done so, that there is something peculiar about technological explanations in that they are intrinsically normative in some sense. Since the notion of artifact function is a normative one (if an artifact has a proper function, it *ought* to behave in specific ways) an explanation of an artifact's function must inherit this normativity.

In this paper I will resist this conclusion by outlining and defending a 'buck-passing account' of the normativity of technological explanations. I will first argue that it is important to distinguish properly between (1) a theory of function ascriptions and (2) an explanation of how a function is realized. The task of the former is to spell out the conditions under which one is justified in ascribing a function to an artifact; the latter should show how the physicochemical make-up of an artifact enables it to fulfill its function. Second, I wish to maintain that a good theory of function ascriptions should account for the normativity of these ascriptions. Provided such a function theory can be formulated — as I think it can — a technological explanation may pass the normativity buck to it. Third, to flesh out these abstract claims, I show how a particular function theory — to wit, the ICE theory by Pieter Vermaas and Wybo Houkes — can be dovetailed smoothly with my own thoughts on technological explanation.

Keywords: technical artifacts, explanation, mechanisms, normativity, proper function.

1. Introduction

To introduce the topic of this paper, here are two observations about technical artifacts. First, technical artifacts have proper functions; that is the very reason behind our designing, making, and using them. They have their functions partly in virtue of their physicochemical make-up. One cannot reasonably ascribe the function to *f* to an artifact, which one knows to have an utterly inappropriate physicochemical constitution — a pencil cannot function as a laptop computer. Hence, there must be some sort of explanatory link between an artifact's function and its physicochemical make-up (or, for short, its 'structure'). When one wants to understand how it is that artifact *x* has the function to *f*, there will be mention of *x*'s structure at some point.

Second, the notion of proper function is a normative one. It makes sense to say of an artifact that it *ought to* exhibit certain behaviors, namely those associated with its function. Such claims do not make sense for normal physical objects, such as stones, solar systems, or sugar molecules.[1] There can be discrepancies between an artifact's proper function and its actual behavioral capacities. An artifact can have the function to *f* even though it cannot *f*. A broken television set still is a television set and the proper function of a worn-out light bulb still is to provide light.[2]

If we combine these two observations we arrive at the conclusion that there is something peculiar about *technological explanations* — i.e. explanations that account for an artifact's function in terms of its structure. Since (1) there must be

[1] At least not in as strong a sense as for artifacts. Of course we can express our (sometimes strongly) inductively supported beliefs about the behavior of physical objects in terms of normative 'ought to'-claims, but it is not as if we have some sort of *right* to expect physical objects to behave as we desire — as *is* the case for technical artifacts (cf. Franssen 2006). More on this in sections 4 and 5.

[2] There are limits here; one would be hard-pressed to still call a television set that has been smashed to a thousand pieces with a jackhammer a television set.

an explanatory link between an artifact's function and its structure, and (2) the notion of artifact function is normative, it seems to follow that technological explanations are special by being inherently normative.

Or so Peter Kroes (1998; 2001) argues. It is my aim in this paper to scrutinize this argument for I think it runs together a couple of different points about artifact functions and explanations. In the next section, I will present Kroes's arguments in more detail. Section 3 contains internal criticism of his arguments, and in section 4 I will argue that there is a more fundamental confusion underlying Kroes's arguments and I will show how we can dispose of this confusion by analyzing his endeavor in two separate projects. We need to distinguish between a theory of artifact functions on the one hand and an account of technological explanations on the other. The former should deal with the normativity of functions, so that the latter can then pass the buck. The rest of the paper serves to flesh out this reply; in section 5 I will present a specific theory of artifact functions and show how it can be combined with an account of technological explanation in the way I envisaged in section 4. Section 6 contains the conclusion.

2. Kroes on Technological Explanations and Normativity[3]

To argue his point about the peculiarity of technological explanations, Kroes first observes that technical artifacts have a dual nature. They are physical objects, but they also have intentional or functional properties essentially. As a result, we can give both functional and physicalistic descriptions of artifacts, with either description partially or wholly black-boxing the other. A clock is any time-keeping device, whatever its exact physicochemical make-up and, alternatively, someone without any experience with pencils cannot deduce that a 6-inch hexagonal elongated piece of wood with a lead inside is for writing (though she might discover that it can be used for writing). The two descriptions are logically independent and, as a result, it is impossible to deduce function from structure or the other way around. Standard deductive-nomological explanations are barred.
Next, he presents an example of a technological explanation that involves the Newcomen steam engine. The main function of Newcomen steam engines was to drive water pumps. They did so by means of the up-and-down movements of

[3] This section summarizes sections 4 and 5 of (Kroes 1998).

their great beam. The great beam itself was driven by the actual steam engine that consisted of a boiler, a steam valve, and a cylinder with moving piston (see Figure 1). Roughly, the explanation of these engines has three ingredients.

(1) Physical laws or phenomena, e.g., that steam occupies a much larger volume than does water, that rapid condensation of steam in a closed vessel creates a partial vacuum, that atmospheric pressure exerts a force on the piston.
(2) The physical make-up and configuration of the engine, e.g., the boiler, steam valve, movable piston, and great beam.
(3) Dynamic behaviors and causal interactions of the components, e.g., heating and expansion of water and steam, opening and closing of the steam valve, injection of cold water, condensation of water, creation of a partial vacuum, and movements of the piston.

Kroes rightly observes that it does not follow from an explanation along these lines that the function of the steam engine is to drive pumps, nor that it is to move the great beam up and down. All that follows is that the steam engine can be used to drive pumps, that it is a means to that end, or that it has the capacity to drive pumps. It is impossible to get the normative *explanandum* containing the ascription of a proper function from the purely descriptive *explanans*. He concludes that the explanation as presented is not a technological explanation since it does not properly account for the steam engine's function in terms of its structure.

Kroes (2001: 38-9) contains a sketchy possible repair. Perhaps, says Kroes, the relation between *explanandum* and *explanans* can be conceived in terms of pragmatic rules of actions that are grounded in causal relations. For example, if one's goal is to drive a water pump, and a steam engine has the capacity to do so (i.e., something like the following causal conditional holds: If the steam engine is put to use properly in appropriate circumstances, it will drive a water pump), then one can infer the following rule of action: To drive a water pump, use a Newcomen steam engine. In this context of action, the steam engine is a means to an end and acquires a function. The engine's physical structure still figures indirectly, since the rule of action is formulated on the basis of a causal conditional that was derived from the engine's structure. Kroes concludes: "A technological explanation, therefore, is not a deductive explanation, but it connects structure and function on the basis of causal relations and pragmatic rules of action based on these causal relations." (2001: 39). So, in sum, Kroes's points are: (1) a technological explanation must account for an artifact's function in terms of its structure, (2) 'standard' explanations (along the lines of the D-N model or a somewhat loosened version of it, as in the example above) cannot accomplish that task, and (3) using the notion of action rules, it seems possible to construe a more adequate account that does connect structure and function in the desired manner.

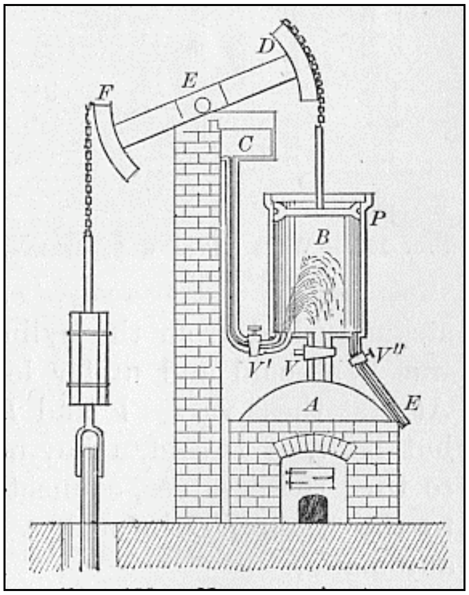

Figure 1 – Newcomen's steam engine

3. Kroes's Arguments Reconsidered

In my opinion, there is something seriously wrong with these arguments. I will argue that Kroes's arguments do not show what they purport to show, even on their own terms, and, in the next section, that his construal of technological explanations runs into trouble because it conflates two rather different projects. As a result, Kroes's effort has to satisfy a set of inconsistent requirements and is doomed to fail.

I can be relatively brief about the first point. It is not clear which of the following claims Kroes aims to establish:

(1) A technological explanation of Newcomen's steam engine does not fit the mold of the D-N model of explanation.

(2) Most or all technological explanations do not fit the mold of the D-N model.

(3) Most or all technological explanations do not fit any of the currently available models of explanation.

I think he should be interested in (3), because that would be a good reason to think that there is something truly peculiar about technological explanations. If the currently available accounts of explanation (such as the D-N account, unification accounts, and causal accounts) are capturing important aspects of what it is to be an explanation, and if technological explanations do not conform to any of these accounts, then they might represent an interesting new species of explanation worthy of philosophical attention. Unfortunately, however, the only claim Kroes establishes with some plausibility is (1). To be fair, I should add that if (1) is correct and the explanation of the steam engine is a representative example of technological explanations in general, the truth of (1) lends inductive support to (2). So to the extent that this inductive argument is compelling, the plausibility of (2) is established as well. But the plausibility of (2) does very little to prove (3). For that, it would have to be shown that technological explanations fit none of the currently available accounts of explanation, e.g. Friedman's and Kitcher's unification accounts, Salmon's, Woodward's, and other causal accounts, Van Fraassen's pragmatic account, and Cartwright's *simulacrum* account. Even accounts of intentional explanation might be relevant if one thinks artifact functions are intrinsically related to agents' intentions. Or accounts of social explanation, if one is of the opinion that artifact functions are inherently social phenomena. I am not saying that this cannot be done, but Kroes has certainly not done it. He has only shown that technological explanations cannot be construed as D-N explanations. While that may be perfectly true, it is hardly a reason for distress, since for many the D-N model has by now been relegated to the domain of philosophical relics. In fact, as I will make clear in due course, there is every reason to think that his construal of technological explanations suffers from internal inconsistencies to such an extent that no account of explanation ought to fit it, on pain of being inconsistent itself.

As far as I can see, the suggestion to construe the relation between *explanandum* and *explanans* in terms of action rules is not successful either. The step from a causal relation to a rule of action is relatively unproblematic: If one wants a certain effect and one knows one or more sufficient cause(s) of this effect, then, given the usual *ceteris paribus* clauses for causal relations and some hedging assumptions about the proportionality of the means in relation to the end, it is perfectly rational from a practical point of view to bring about this effect by bringing about one of these sufficient cause(s). If Newcomen's engine can drive a water pump if it is operated properly, one could use it to pump water if one wants so, but — and this illustrates the chief difficulty — in the same vein we can add that, if it can be used to tear stuff apart, one could use it to tear stuff apart if that is what one wants. One can use an electric guitar to play licks, and if one so desires, it would be rational to use it for that purpose, but if one is in a rockstar-type of mood, a guitar can also be used to smash loudspeakers, and it would be no less rational to use it to that end. None of this, however, goes to show that Newcomen's steam engine is for tearing stuff apart or that smashing loudspeakers is an electric guitar's proper function.

Although the fact that something has a number of capacities that can be expressed in terms of causal conditionals warrants inferences to various rules of action (under the assumptions mentioned), nothing supports one of these rules in particular as the *proper* one, and neither does the artifact considered in isolation give you any reason to suppose that one of these causal capacities is the artifact's *proper* function, as opposed to an accidental or system function, i.e. just something it can do. While causal knowledge may underpin rules of action, I do not see how it could sustain proper function ascriptions. In the end, the suggested repair is not much of an improvement over Kroes's initial proposal. All that can be inferred from action rules is that if a certain artifact can be used to accomplish some end, then it is rational to use it to that end, but that follows virtually analytically (again, given some background assumptions) from the fact that it *is* a means to that end, and that was already established in the initial proposal.

4. Functions: To Ascribe and to Explain

Given that Kroes's project leads to a dead end considered by its own lights, let us now take a step back and turn to the second point. I will argue that there is a more fundamental confusion vexing the project. Unearthing this confusion will also enable us to see why his project really was a non-starter. Kroes stipulates that a technological explanation is an explanation that accounts for an artifact's proper function in terms of its physicochemical make-up. This construal is, I think, seriously misguided because it runs together two rather different projects, to wit (1) that of giving an account of proper function ascriptions and (2) that of explaining how, in virtue of its physicochemical make-up, an artifact can fulfill its function. The result of (1) is a set of necessary and jointly sufficient conditions for the truth or assertibility of claims like 'artifact *x* has proper function *f*.' It is fairly obvious that this set will contain more conditions than just those related to the *x*'s physicochemical make-up — that is in fact the negative result of Kroes's argument: claims about proper functions cannot be deduced solely from information about the artifact's physicochemical make-up. But it is not so obvious that something like a highly detailed account of *x*'s workings must be among these conditions, for that would mean that no one except highly knowledgeable engineers could ever be justified or correct in claiming that an artifact has a proper function. Project (2), on the other hand, provides an account of how an artifact's physicochemical make-up enables it to exhibit the behaviors required for its proper functioning and here the notion of mechanistic explanation immediately springs to mind. What Kroes tries to do, however, is to get the results of both project (1) and (2) while drawing exclusively on the means for project (2). That is an impossible task.

An analogy will help to clarify the reason why. Suppose we want to explain why the function of the heart is to pump blood, or, more precisely, to determine whether the proper function of the heart is to pump blood. Surely, an elaborate scrutiny of hearts and their behavior by itself will not allow us to conclude that their proper function is to pump blood, yet this is the only option open to us on an extrapolated version of Kroes's proposal, since he seems to be thinking that an item's proper function could be determined just by looking at its physicochemical make-up. Instead, we should distinguish the project of spelling out the truth or assertibility conditions for "The function of the heart is to pump blood", from that of explaining how the heart is able to pump blood. Accounting for the fact *that* the function of the heart is to pump blood is not the same as accounting for *how* it can pump blood. The first project will involve more than just the heart's 'intrinsic' properties. Biological function theories disagree on exactly what more; some suggest synchronic relational properties such as the heart's current contribution to organism fitness (Walsh 1996; Lewens 2004), others look at diachronic relational (historical) properties such as the heart's ancestors contribution to ancestor fitness (Millikan 1984, 1993; Neander 1991a, 1991b). The outcome of the second project, however, will look more like Kroes's proposed *explanans*. It will explicate how the physicochemical make-up of the heart and its constituent parts in their particular configuration leads to dynamic behaviors that, in the appropriate environment, add up to pumping blood.

The crucial point is that accounting for an item's proper function, on the one hand, cannot be done without taking the item's environment into account, be it its current ecological niche, its history, its ancestors, its users, its designers, or their intentions and/or (justified) beliefs. Proper functions are not among the intrinsic properties of an item and therefore they cannot be discovered by solely looking at the item itself, isolated from its environment. An item's capacities and its behaviors, on the other hand, are among its intrinsic properties and can be explained by looking just at the item's physicochemical constitution and mereological make-up. The two projects are largely independent. One can be justified, even correct, in ascribing proper functions to organs or artifacts without knowing how they are able to perform that function, and, alternatively, one can explain how it is that organs or artifacts (or their parts) have the capacities they have or show the behaviors they show without knowing that one of these capacities or behaviors is associated with a proper function. Of course, one is typically interested in an explanation of how an organ or artifact can perform the behavior associated with its proper function, since that tends to be its most interesting feature (that computers can function as paperweights is not the reason people buy them).

What I have said so far should not be taken to imply that the projects are entirely unrelated; I have only argued that it is unwise to try and tackle them in one fell swoop. I now want to look at possible connections, two in particular. The first one is that an explanation of how something is able to perform its function might pop up in the justification for its having that function. Roughly, the intuition is that function ascriptions must have something to do with the actual behavioral capacities an object has, at least for paradigm exemplars of the object. In order to justify the claim '*x* has proper function *f*' (where *x* is a normal exemplar of its type) there must be evidence that *x* can in fact *f*, and an adequate explanation of how *x* can *f* would be very good evidence, albeit not the only permissible type of evidence. Naïve theories of artifact functions overlook this intuition. Consider a theory that defines the function of an artifact to be what the designer intended the artifact to do. Such a theory lacks the evidence-requirement and thereby fails to link claims about proper functions to (evidence of) actual capacities. As a result, it allows for crazy function ascriptions. A mad designer's *intention* to build a spacecraft from a bunch of matchsticks does not warrant the conclusion that the result he produces *is* a spacecraft, for there is no way in which matchsticks could ever compose a spacecraft, at least not by current scientific lights. So the first way in which the two projects are related is by way of justification. An explanation of how something can fulfill its function can be among the justificatory grounds for the claim that an artifact has that proper function.

The second connection appears in malfunction cases: situations where an item still *has* a proper function, even though it cannot *perform* that function. I assume that such cases do exist, both in biology and technology; malfunctioning hearts are still for pumping blood, and the proper function of a worn-out light bulb still is to provide light.[4] For malfunction cases, the second project I identified takes on a slightly different form, since the question of how the artifact can perform its function is obsolete when we know that it cannot perform its function. What can be explained, however, and what is not obsolete, is how the artifact was *supposed* to perform its function. An answer to that question will look a lot like the answer to the original explanatory question, except that it will be phrased in normative or counterfactual terms. It explicates how the various parts *ought* to be configured, behave, and interact, or how they would have been configured and how they would have behaved and interacted, were the artifact to function properly.[5] Even if one does not think that this answer is valuable in and of itself, it should be obvious that it has instrumental value as background knowledge for determining the causes of malfunction. Only in contrast to how the artifact was supposed to work will it become possible to find out how it malfunctions.

[4] One might argue over whether cases of worn-out artifacts properly belong under the heading of malfunction. For example, light bulbs are apparently designed so as to stop working after a certain amount of burning hours. I can see that one might interpret this as evidence that wearing out is in fact part of the proper function of a light bulb. For brevity’s sake I will ignore this terminological quibble while taking it to be uncontroversial that a worn-out artifact still has a proper function.

[5] Establishing the truth of counterfactual claims is a notoriously troublesome issue, which I cannot hope to address to any satisfactory extent here. I rely on an intuitive way of thinking about it, but will add one important qualification. The possible worlds taken into account must be limited to those close to our own with roughly similar laws of nature. Without this constraint, it may be possible to think of worlds where materials and artifacts have very different properties and capacities so that, say, a bunch of matchsticks could compose a spacecraft. If this brief remark does not satisfy the reader, my advice is to forget about the counterfactual reading altogether and focus on the normative reading.

Unlike scientific explanations of natural phenomena, technological explanations can inherit the normativity of function ascriptions. Although we might claim that photons 'ought' to behave as particles, this only goes so far as the theory from which we infer this claim has been inductively supported or as our previous experiences lend inductive support to such a claim. Such claims merely express inductively supported expectations about phenomena. Technological explanations, however, can incur an extra and stronger type of normativity in that there are independently ascertainable and objective facts of the matter as to how the artifact and its components ought to behave. These facts are grounded in the justified beliefs, intentions, and communicative actions of the designer(s) who devised the artifact or in the beliefs, intentions, and actions of the (group of) users who put the artifact to a new use that has gradually become widespread standard use.[6] Under the assumption that she is *competent*, i.e. broadly rational and skillful and in possession of appropriate justification for the beliefs upon which she acts, a designer objectively determines an artifact's proper function. That fact entitles us to objective claims about what this proper function is, even in the face of malfunction. Note that the competence assumption is essential: only if designers tend to have correct beliefs about the workings of the artifacts they devise, skillfully build the artifacts they devise (or see to it that this gets done), and truthfully communicate about functions, will we have additional reasons, beyond mere past experience or other inductive support, for claiming that artifacts ought to behave such-and-such.

Looking at the kinds of justification involved can further bring out the difference. The justification for ought-claims about malfunctioning artifacts differs in kind from the sorts of justification we might have for normative statements about the behavior of natural objects. Of course, designers base their beliefs on scientific theories or practical experience with the materials they use and in this sense their knowledge about artifacts parallels the sort of knowledge scientists have about natural phenomena. An engineer's claim that an iron bar ought not to buckle under a specific pressure does not differ in kind from the claim that a photon ought to behave as a particle; both are supported by normal scientific evidence. For non-designers, however, another story must be told. Provided the

[6] For brevity's sake, I will ignore such user-imparted proper functions for the rest of this section, but a story very similar to the story I am about to tell can be told about them.

competence assumption mentioned above is warranted — as it certainly seems to be in our society — they can take the designer's word[7] as support for claims about proper functions and hence about how the artifact and its components ought to behave. What is more, because of social, economical, and legal arrangements in our society, users have legal rights vis-à-vis designers with regard to claims about what artifacts ought to do. Warrant for the competence assumption is officially institutionalized, so to speak. Designers are expected to be trustworthy and reliable in what they do, i.e. they are expected to be competent. Failing these expectations leads to sanctions. Because of all this, non-designers are entitled to objective normative claims about the proper functions of artifacts. The justification for such claims consists of beliefs about what the designers wanted an artifact to do. These beliefs screen off other types of justification, such as experience with the artifact, testimony about successful artifact use, or even a theory about the artifact. Of course, non-designers sometimes also have these latter justifications for a claim that artifact *x* ought to *f*, but my point is that they do not *need* it in order to be justified in claiming so. All they need for that — still under the competence assumption — is knowledge or justified belief that the designer intended *x* to be for *f*-ing. This screens off other types of justification and provides just the extra normative force that adheres to proper function claims and that can be inherited by explanations of how a malfunctioning artifact ought to function.

To round off this lengthy excursion about the normativity of proper function claims, let me give an illustration. Say I have been commuting happily in my car every day for the past year, but then one morning when I turn the key it will not start. I want to claim that my car still has its proper function (say, personal motorized transportation) and that it ought to start if I turn the key, even though it presently malfunctions. What sort of evidence do I have for this claim? Obviously my past experience with the car, but that is not crucial. What is more important is that I have every reason to believe that my car was designed and built by competent engineers with the purpose of designing and building something that has the proper function of providing personal motorized transportation. Therefore, the normative claim that my car ought to start carries with it an extra and stronger normative force beyond that offered by the mere past experience induction. If I were to not have had that experience, I would still

[7] Or something derived from that through a chain of communication, e.g. what the label on the box says, or the salesperson, or your sister who just bought the artifact.

have been entitled to claim that my car ought to start. Compare this with my successfully and regularly lulling a child to sleep by the monotonous sound of driving. If one day the child will not go to sleep, I might also claim that my car ought to lull the child into sleep, but this claim clearly carries a smaller normative force because it lacks the additional support of claims related to proper functions.

What does all this mean for the (alleged) inherent normativity of technological explanations? Let me spell out the ramifications of what I have said.

(1) Proper function ascriptions have a normative force in that they can be correct of an artifact even when that artifact cannot perform its proper function. Malfunctioning artifacts still have proper functions.[8]

(2) An adequate theory of artifact functions — the result of what I dubbed project (1) — should account for this normativity, i.e. it should reproduce the better part of our intuitions about which malfunctioning artifacts nonetheless have proper functions.

(3) An account of technological explanation — the result of project (2) — may pass the normativity buck to the theory of artifact functions and need not account for normativity itself.

(4) Technological explanations can inherit the normativity of function ascriptions. If an artifact functions properly, a good technological explanation truthfully explains how it does so. If an artifact malfunctions and still has a proper function, a good technological explanation explains — in equally truthful ought-claims or counterfactuals — how the artifact was supposed to function, had it not been malfunctioning.

The obvious question is whether the two projects I have outlined are feasible. It is one thing to formulate a set of requirements that a theory of functions and an account of explanation ought to satisfy, but quite another thing to show that these requirements can be met. That is why I will use the next section to sketch a theory of functions and an account of technological explanation — the former borrowed, the latter of my own making — that, for all I can see, satisfy the requirements I have submitted.

[8] One might argue over whether the notion of malfunction presupposes that of a proper function so that every malfunctioning artifact necessarily has a proper function. Nothing much depends on this for me, so I leave the question undecided.

5. Making It Work

Wybo Houkes and Pieter Vermaas have developed a theory of artifact functions which seems to me perfectly suitable for the present purposes (Houkes and Vermaas 2004; Vermaas and Houkes 2006). First, a bit of background. On this theory, artifacts are embedded in the action-theoretical notion of a *use plan*: a series of considered actions undertaken to realize a practical goal desired by an agent, in which at least one of the actions involves the manipulation of the artifact. By exercising one or more of its capacities an artifact contributes to the realization of the overall goal of the plan. Designing engineers devise use plans when they design artifacts, but users are free to invent their own alternative use plans, which may subsequently become new standardized uses. The theory itself, then, is a theory about when agents are justified in ascribing functions to artifacts. Here is what it says.

> An agent *a* [justifiably, JdR] ascribes the capacity to *f* as a function to an artifact *x*, relative to a use plan *p* for *x* and relative to an account *A*, iff:
>
> A. the agent *a* has the belief that *x* has the capacity to *f*, when manipulated in the execution of *p*, and the agent *a* has the belief that if this execution of *p* leads successfully to its goals, this success is due, in part, to *x*'s capacity to *f*;
>
> B. the agent *a* can justify these two beliefs on the basis of *A*; and
>
> C. the agents *d* who developed *p* have intentionally selected *x* for the capacity to *f* and have intentionally communicated *p* to other agents *u*.
>
> (Houkes and Vermaas 2004: 65, with slight notational adjustments)

A few remarks for clarification. First, on this account functions are relativized to use plans; the latter is the more fundamental notion. Having a function means for an artifact to be embedded in a use plan that privileges one (or a few) of its many capacities as special, i.e. as its proper function(s). Secondly, the beliefs that *x* can *f* and that its doing so contributes to the realization of the use plan's goal need to be justifiable on the basis of an account *A* (which is itself subject to normal standards of justification). This account can take on a number of forms; for new, inexperienced users it can be simple testimony or observation (having heard that this contraption is a laser pointer, or having read the inscriptions on the package), for technically savvy users who enjoy taking apart their electrical appliances, it can be practical insight in their internal workings combined with experiential knowledge, and for engineers, it will typically be full-fledged technological and scientific explanations, often combined with practical experience from prototype tests. Thirdly, as foreshadowed in the previous sections, the notion of function turns out to be a relational one. To put it somewhat crudely, artifacts by themselves do not have functions; they acquire functions in a context of use plans, users and designers, and their justified beliefs, intentions, and actions.

Does this function theory account for the normativity of proper function ascriptions? Its creators think it does and I am inclined to agree with them. For brevity's sake, I will not laboriously go over a host of examples that the theory successfully covers, but limit myself to an outline of its general strategy for coping with the normativity of function ascriptions and a discussion of one worry.[9] Since agents only need justified beliefs, as opposed to knowledge, about the artifact's capacities, the theory allows for cases in which an agent's beliefs are defeated by later evidence. In this way, one can ascribe functions to malfunctioning artifacts. I may have every reason for believing that my phone has the appropriate capacities to allow me to call my mother and fulfill all the other conditions laid down by the theory and, by that token, be justified in ascribing the function of allowing for conversations at a distance to it, but if — unbeknownst to me — a practical joker has removed the microphone from my phone, it will nonetheless malfunction.

This example, however, does raise a concern, for the theory seems to imply that once I have learned of my phone's malfunctioning, I can no longer ascribe the same proper function to it because I no longer have the belief that it has the capacity to transmit my voice to the other end of the line. That is a counterintuitive result. To deal with cases like these, we must modify condition I. The agent does not have to have the belief that *x* can *f* but may also have the overriding belief that *x* would have been able to *f*, had particular counteracting interferences not occurred, or that it ought to be able to *f* given what the designers communicated about *x*. In short, condition I should read: agent *a* has the belief that *x* has or *should* have the capacity to *f*, etc. (and, of course, *a* must be able to justify this belief too). With this modified condition in place, I can still ascribe the function of teleconversation to my phone after learning about the removed microphone, for I am justified in believing that it would have had that capacity, had someone not been playing this joke on me.[10, 11]

[9] A more elaborate discussion of the theory can be found in (Houkes and Vermaas 2004, 2005; Vermaas and Houkes 2006).

[10] To be complete, I should add that for situations where an artifact malfunctions due to normal wear and tear, the I-condition must be modified to include something like 'agent *a* knows that *x*

So far so good then. The next task is to see if this function theory matches up with an account of technological explanation in the way I envisaged. Not unexpectedly, I think it does. As I argued above, such an account of explanation must deal with explanations that explicate how artifacts are able to exhibit various behaviors, and the behavior associated with their function in particular. I think the resources for this are available in the literature on mechanistic explanation, although they have not always been clearly recognized and presented. I have given the contours of this account of explanation in another paper (De Ridder 2006) and I will briefly summarize my ideas here. To explain a particular piece of artifact behavior, there are two general strategies available, leading to two different complementary types of understanding (cf. also Bechtel and Richardson 1993: 18). I have tried to capture these strategies in the following descriptions.

> *Top-down strategy*: take the behavior to be explained and decompose it into more basic sub-behaviors, reiterate this step if possible — it should become clear how the complex behavior being explained is realized by simpler behaviors in a specific spatiotemporal configuration — and for all the sub-behaviors, indicate which component(s) take(s) care of them.
> *Bottom-up strategy*: identify the structural components of the artifact and give information about their physicochemical make-up and spatial configuration, show how their physicochemical features and configuration result in various behaviors and then describe how these behaviors, in their spatiotemporal configuration, together make up the behavior to be explained.

The first strategy focuses on behaviors; it explicates how a complex behavior is realized by ever-simpler sub-behaviors by decomposing the overall complex behavior in its constituent sub-behaviors. It provides purely *functional*[12] understanding, solely in terms of behaviors, thereby black-boxing the physicochemical make-up of the artifact and components that exhibit these behaviors. The second strategy opens up the black box; it starts from the

used to have the capacity to *f* but has now stopped having that capacity due to normal wear and tear'.

[11] The suggested modifications are in line with what Houkes and Vermaas say, but not explicitly theirs.

[12] Note that, in this context, the term 'functional' has the weak sense of 'having to do with input-output relations only'; it does not refer to the richer notion of proper function.

structural decomposition of the artifact by identifying its component parts, and then describes their relevant characteristics (morphological, physical, and chemical properties relevant for the behavior being explained), and how these characteristics enable particular behaviors (under appropriate circumstances). Finally — and here it overlaps the first strategy — it shows how these behaviors add up to the complex behavior being explained. The second strategy offers *structural* understanding of the artifact's workings. So while the first strategy starts from a decomposition of the behavior and subsequently indicates how the structural parts fit into this functional, or behavioral, decomposition, the second strategy starts from the structural decomposition and works its way upwards to the behaviors exhibited by the structural components, showing how the behavioral decomposition maps unto the structural decomposition. Although the two strategies are complementary I do not think they should be merged into one. The demands that this merged strategy would place on a good technological explanation are too strict. Explanations that only provide functional understanding would automatically be disqualified as incomplete, whereas I am convinced — although I will not argue the point here — that they are perfectly good explanations in many contexts, not just in the pragmatic sense of being acceptable to the person with the explanatory request, but also in the stronger sense of being an objectively good explanation.

I hope this brief sketch suffices to give the reader an impression of what this account of technological explanation looks like. Let us now move on to the last part of the paper and see if this account lives up to the standards I set for it in the previous section. The crucial question is whether it can grapple with the normativity issue for malfunction cases. If we ascribe a proper function to a malfunctioning artifact on the basis of the modified ICE conditions there are three options: (1) we have a false but justified belief that the artifact has the capacity to function, (2) we have a justified (and true) belief that the artifact should have the capacity to function (in the sense described earlier), or (3) we know that the artifact used to have the capacity to function, but that it is now worn-out. In all three cases, it seems to me perfectly possible to give an explanation of how the artifact is believed to work, supposed to work or how it used to work. In each case and for both explanatory strategies, the explanans will be phrased in normative or counterfactual terms, e.g., this component should sit here and interact with that other one right there so that they would have shown such-and-so behavior, thus contributing to the proper functioning of the artifact. Or: this light ought to go on when I hit that switch. Or: this spring used to push that thing back. Like explanations of properly functioning artifacts, these explanations can be evaluated in terms of truth, justification, acceptability, or whatever else is deemed appropriate. I don't see any particular problems about normativity left here that an account of technological explanation could not pass on to a function theory.

Someone may worry about circularity, though. If part of the justification for ascribing a function is a technological explanation and if an account of technological explanation passes the normativity buck to the function theory, doesn't that land us in some sort of justificatory circle? Not if we look closer at the exact justificatory relations. Typically, professional engineers or technically savvy 'laypersons' will have at their disposal more or less elaborate technological explanations as justifications for (some of) the function ascriptions they make. That means that they will have justified beliefs about the physicochemical properties of the artifact and its components, the components' configuration and interactions, and their behavioral capacities. But the justification for these beliefs, and hence for the explanation, in no way depends on the function ascription; instead it is based on the normal justificatory mechanisms for beliefs about stuff in the world: observation, experiments, experience, and testimony. So if an engineer ascribes a function to a malfunctioning artifact, the normativity of this ascription is in the end epistemic, derivative of the normativity of epistemic justification. Although the justification for a function ascription will, for some persons, rely on a technological explanation, the justification for this explanation in its turn does not rely on the function ascription and therefore there is no justificatory circularity here.

Persons lacking access to technological explanations who make function ascriptions justify these ascriptions by observation, experience, or testimony. In addition to the normative force of good justifications, these laypersons have an additional normative claim that entitles them to say that an artifact ought to have a certain proper function and fulfill it properly, as elaborated in the previous section. The epistemic division of labor in our society is such that professional engineers are entrusted with the task of designing properly functioning contraptions for various purposes. Laypersons have a legal and 'social-epistemic' right to expect engineers to have true beliefs about the workings and functions of the artifacts they make and to trust their testimony.[13] Whatever the details of this arrangement, we do not stumble on a justificatory circle here and that is the point

[13] I owe this point to Wybo Houkes, cf. also (Houkes and Vermaas 2005) esp. chapter 8.

I wanted to argue. The circularity worry is misplaced and the combination of the ICE theory of function ascriptions and my account of technological explanation can bear the burden. For all I can see, the two together deal adequately with the normativity of proper function ascriptions and technological explanations.

6. Conclusion

Against Peter Kroes I have argued that technological explanations are not necessarily special because they have to deal with the normativity of proper function ascriptions. Kroes's argument rests on a confusion of two rather different projects: that of giving a function theory and that of giving an account of technological explanation. The first project should grapple with the normativity of function ascriptions, i.e. it should explicate the conditions under which malfunctioning artifacts have proper functions. The second project can then pass the normativity buck to the first project. The principal reason for distinguishing these projects is that the property of having a proper function is relational, or extrinsic, whereas the property of having the capacity to exhibit a particular behavior is intrinsic. Consequently, accounting for the property of having a proper function must take the artifact's context into account, while accounting for the property of having a behavioral capacity can be done by looking just at the artifact itself. I have also argued that my way of framing the problem is more than wishful thinking, because Vermaas and Houkes's ICE function theory and my account of technological explanation do a good job in meeting the requirements I set out for the two projects. Besides, they fit together fine in the way I envisioned at the beginning of this paper.

References

Bechtel, William, and Robert C. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton, NJ: Princeton University Press.

De Ridder, Jeroen. 2006. Mechanistic artefact explanation. *Studies in History and Philosophy of Science* (forthcoming).

Franssen, Maarten. 2006. The normativity of artefacts. *Studies in History and Philosophy of Science* (forthcoming).

Houkes, Wybo N., and Pieter E. Vermaas. 2004. Actions versus functions: A plea for an alternative metaphysics of artifacts. *The Monist* 87 (1): 52-71.

———. 2005. *Artefacts: From Functions to Plans of Use*. Book manuscript.

Kroes, Peter A. 1998. Technological explanations: The relation between structure and function of technological objects. *Technè* 3 (3): 18-34.

———. 2001. Engineering design and the empirical turn in the philosophy of technology. In *The Empirical Turn in the Philosophy of Technology*, edited by P. A. Kroes and A. W. M. Meijers. Amsterdam: JAI (Elsevier Science).

Lewens, Tim. 2004. *Organisms and Artifacts: Design in Nature and Elsewhere*. Cambridge, MA: MIT Press.

Millikan, Ruth. 1984. *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

———. 1993. *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.

Neander, Karen. 1991a. Functions as selected effects: the conceptual analyst's defense. *Philosophy of Science* 58 (2): 168-184.

———. 1991b. The teleological notion of 'function'. *Australasian Journal of Philosophy* 69 (4): 454-468.

Vermaas, Pieter E., and Wybo N. Houkes. 2006. Technical functions: A drawbridge between the intentional and structural natures of technical artefacts. *Studies in History and Philosophy of Science* (forthcoming).

Walsh, Denis M. 1996. Fitness and function. *British Journal for the Philosophy of Science* 47 (4): 553-574.

How Norms In Technology Ought To Be Interpreted

Krist Vaesen
Technical University of Eindhoven
Philosophy and Ethics of Technology

Abstract: This paper defends the claim that there are — at least — two kinds of normativity in technological practice. The first concerns what engineers ought to do and the second concerns normative statements about artifacts. The claim is controversial, since the standard approach to normativity, namely normative realism, actually denies artifacts any kind of normativity; according to the normative realist, normativity applies exclusively to human agents. In other words, normative realists hold that only "human agent normativity" is a genuine form of normativity.

I will argue that normative realism is mistaken on this point. I will mainly draw on material of Daniel Dennett and Philip Pettit to show that it makes sense to talk about artifactual normativity. We claim that this approach can also make sense of human agent normativity — or more specifically "engineer normativity". Moreover, it avoids some of the problems formulated by opponents of normative realism. Thus I will develop a strategy which: (i) makes sense of artifactual normativity; and (ii) makes sense of "human agent normativity", specifically "engineer normativity".

KEYWORDS: Normative Realism — Response-dependence — Normativity — Technology — Interpretation

1. Introduction

In a now classic paper Hector-Neri Castañeda developed a theory of normativity consisting of two main categories: the category of ought to do and the category of ought to be [1]. Some authors accepted this distinction, while offering more elegant formulations. They would, for example, rather

[1] Castañeda, 1970.

talk about deontic normativity (ought to do) and evaluative normativity (ought to be). Others have doubted such a distinction can be made at all. Roughly stated, they claim that the evaluative can be reduced to the deontic. For instance, to say an act was right, means nothing more than that the agent has done what he ought to have done [2].

At first sight, it seems the philosophy of technology could benefit from Castañeda's distinction, since it seems apt to define two forms of normativity in technology. The category of ought to do would in that case cover statements about what the engineer ought to do — how he ought to design his artifacts, for instance. The category of ought to be, on the other hand, would relate to how artifacts ought to be — e.g. the dimension of a piece of A4 paper *ought to be* 210mm x 297mm.

To some extent I will defend and make more explicit this line of argument in the course of the present paper. However, I will formulate it differently. From now on, I will not speak about "ought to do" and "ought to be", but about "human agent normativity" and "artifactual normativity". On this account artifactual normativity not only comprises ought to be statements, but some ought to do statements as well, such as: this artifact is a watch, so it ought to *perform* its intended function, namely it ought to *keep* the time.

The aim of this paper, then, is to answer the question: can we make sense of this so-called artifactual normativity? I will contend that, indeed, we can.

I will proceed as follows. First, I will say something about normative realism, which is arguably the standard approach to normativity. I will argue that normative realists cannot account for artifactual normativity. They may argue that this is no problem at all, since there is no such thing as artifactual normativity. I will argue that this artifactual normativity is, on the contrary, essential for making sense of engineering norms.

Indeed, I will go a step further. I will argue that normative realism is a poor candidate to account for human agent normativity as well. To do that, I will invoke two points of criticism, which I call the *problem of autonomy* and the *problem of intentionality*.

In sections 4 and 5, I will develop an alternative that makes sense of artifactual normativity and, at the same time, makes better sense of human agent normativity. It will be interpretative and dispositional in nature.

[2] Dancy, personal communication.

In section 6, I will programmatically deal with some ontological issues and suggest that my account is not anti-realist nor relativist. Finally, section 7 ends with some concluding remarks.

2. What Normative Realists Ought To Reconsider: Part I

Clearly, I have some doubts about what normative realism as a theory might achieve, but let us begin with a short summary of the theory.

Normative realists maintain that normativity can be explained — if it can be explained at all — in terms of reasons. More importantly, normative realists think those reasons are facts, facts which, more or less independently of our human make-up, provide reasons *in virtue of their own nature* [3]. So, if Jesse has a reason not to play with guns, the normative realist would say, it is because what playing with guns consists in, and not because of Jesse's psychological make-up, desires, and the like. Playing with guns is objectively wrong and this fact gives people a good reason not to play with guns.

In the 1970's, Joseph Raz first explored normativity in terms of reasons. He remains loyal to the basic idea: 'The normativity of all that is normative consists in the way it is, or provides, or is otherwise related to reasons.' [4]

Raz characterizes a norm as follows [5]: A norm is a fact which operates as an *exclusionary* reason. This means that a norm contains not only a first-order reason, stating what one ought to do, but a second-order reason as well, namely an exclusionary one: it excludes acting for another (competing) reason. Consider for instance the imperative "Obey your superior". According to Raz, it derives its normativity from the fact that it provides a reason to obey your superior and that it implies furthermore that you ought not to act for any other reason; irrespective of other considerations, you ought to obey your superior.

Following von Wright, Raz discerns four elements in any norm [6]: (i) the *deontic* operator (the so-called ought); (ii) the norm subjects, namely the persons required to behave in a certain way; (iii) the norm act, namely the action which is required of them; and (iv) the conditions of application,

[3] Here I follow a formulation of Lillehammer (2002 and 2003).
[4] Raz, 1999, p. 67.
[5] Raz, 1975 & 1990.
[6] Ibidem, p. 50.

namely the circumstances in which they are required to perform the norm action.

For my purpose, element (ii) is crucial: according to Raz norms only apply to human agents and their actions. Human beings might have reasons to do such-and-such, objects don't. Corollary: unless Raz agrees to take a Dennettian interpretative stance towards artifacts — and I am confident that he, as a realist, would not — artifactual normativity does not fit into his normative realist account. I will later argue that this is a deficiency, but let us first show that Raz is not alone in this conclusion. We will examine a second normative realist, Jonathan Dancy, and show that he is committed to the same conclusion — artifactual normativity is unexplainable for the normative realist. Along the way, I will sketch some arguments to be used in later sections.

Dancy's particularism reacts to the fact that authors, such as Raz, restrict normativity to so-called perfect reasons, i.e. reasons which cannot be overruled by any other consideration [7]. Dancy claims that such perfect reasons don't exist. He defends the thesis that all reasons are *pro tanto* [8]. Consider the norm "Obey your superior". In some cases, it indeed excludes disobedience, in other cases, the norm will be weighed against other reasons; your superior might be drunk or may ask you to do something highly despicable. A *pro tanto* reason, then, is a reason which is, all things considered, the best reason to act upon.

If we agree with the particularist, we can not understand the normative merely in terms of reasons which *exclude* other options — like Joseph Raz does. Dancy's account attempts to reconcile normative realism and particularism by explaining normativity in terms of "favouring". He holds that normative reasons are reasons which favour certain paths of action and, importantly, that favouring comes in different degrees. Some facts are more decisive than others; some speak modestly in favour of doing X, others cry out loud, so to speak. Confronted with conflicting reasons, agents weigh them and generally select the most favouring reason for action. All things considered, it is the best reason at hand.

[7] More accurately, Raz allows for non-perfect reasons as well. He calls them non-mandatory reasons. This means they are mildly exclusionary; they permit you to refrain from other (competing) reasons. It is however hard to see how such non-mandatory reasons differ from ordinary reasons.

8 Dancy, 2004.

Thus, like Raz, Dancy reduces normativity to reasons for human action [9]. Maintaining that a knife ought to be sharp, for example, is a non-normative claim, unless it favours further action, such as giving the engineer a reason to produce sharp knifes.

I hope it's clear that both in Raz's and Dancy's account there is no room for artifactual normativity. But why should this be a problem?

First, it is widely agreed that normativity is related to certain ought statements. The exact nature of this relationship and whether oughts can be analyzed at a more primitive level remain open issues — or, better, topics for philosophical dispute. Saying, as normative realists do, that normativity is related to reasons is one thing; defining, for instance, when mere reasons turn into normative reasons — presumably *the* crux of the question: what is normativity? — is another. And so far — as Jonathan Dancy concedes[10] — nobody has come up with a satisfying answer to this question, except by relying on mere intuitions. Pending settlement of this point, I think it is reasonable to maintain a rather liberal account of normativity and claim that my intuitions are different from those of the normative realist: ought statements about artifacts are normative. Of course, this is too easy a way out, so I will develop two other lines of argument.

First, restricting normativity to human agent normativity seems in conflict with our everyday use of the term "norm". Consider an example taken from the technological sciences: the Dutch Institute for Norms and Normalization. It is, so to speak, a gathering point for norms; norms which not only concern the conduct of engineers producing artifacts, but also the artifacts themselves. For a car to be marketable it ought to function properly *and* it ought to conform to certain standards – for instance, it ought to comply with certain emission standards, it ought to pass such-and-such crash tests, and the like. Of course, such norms might function as a motivational element in human behavior, for instance, in the behavior of engineers designing the artifacts which ought to be so-and-so. But such norms seem independently relevant in legislation and in cases where a user evaluates a certain artifact. In such cases, the primary focus of the evaluator is the artifact, not the behavior of its designer. The most natural way to describe norms, I suggest, is to think of them as idealizations of how things ought to be done or ought to be. For example: in order for a human agent to meet the norms of rationality, he ought to act so-and-so; in order for an artifact to meet the norms of

[9] Dancy, personal communication.

[10] Dancy, personal communication.

optimality, it ought to be so-and-so, or ought to perform this-and-that. On this account, then, the difference between agent and artifactual normativity would be related to what the norms are *about*: in agent normativity, norms are about human beings and their actions, artifactual normativity on the other hand concerns artifacts.

Of course, the normative realist could grant that in natural language we do indeed use normative notions when talking about artifacts; but, as a philosopher, (s)he has taken up the job to tidy up the sloppiness of natural language. (S)he might do this by following at least two other lines of argument. The first is to claim that so-called artifactual norms are non-normative, since they merely refer to expectations. Second, the normative realist could hold that these so-called norms are to a certain extent normative, but in a derivative sense: their normativity ultimately can be reduced to norms about actions, say, the designer ought to have taken. By giving three examples, I will show the problems of both strategies, and thus put the burden of proof on those who deny artifacts any kind of normativity.

First, if artifactual norms were expressions of mere expectations, it would be hard to understand cases in which expectations yield evaluative judgments. Suppose I drop a pen. My expectation is that it will fall. This is sometimes expressed as, "when I drop it, it *ought* to fall," but clearly the ought here is non-normative. It does not support evaluative judgments: if the pen somehow fails to fall, I wouldn't judge it a bad pen. Nor will I call the manufacturer to tell him the pen was poorly designed. On the other hand, if I use it to write down something and see that no ink is released, my claim 'The pen ought to release ink', is not only about what I expect the pen to do, but also relates to what it (normatively) *ought* to do, *given its intended function*. Only when I have such intended function in mind, I am in a position to judge the pen to be 'good' or 'bad', a judgment I wouldn't make about the pen disobeying the laws of gravity.

The example illustrates that artifactual oughts may have two sides: an evaluative one *and* one casting expectations. A second example, however, can show that some of those oughts are merely evaluative, and even stronger, that they are at odds with our expectations. Suppose you find your car, lights still on. They still glow, but only dimly; presumably your battery has run low. In this case, your judgment 'my car ought to start', surely doesn't reflect what you expect: you reasonably believe your battery has run low, so you predict that your car will *not* start. Again, your judgment is framing what the car ought to do, in order for it to fulfill its functional role. You may take yourself responsible for the car's malfunctioning, or shift responsibility to its producer: the latter has done a poor job, (s)he should have built in an automatic light extinguisher.

A final example should make clear why evaluative judgments about artifacts not always can be reduced to agent normativity. Indeed, if your new car is malfunctioning, you can hold its producer responsible: (s)he ought to have designed it so that it, say, does not explode when you turn the ignition on. But what if your car is malfunctioning just because it's an old one? Because it *is* a car, there are certain functions it ought to perform: for instance, it ought to start when I turn the key. Since it doesn't, it is a malfunctioning car: it's a poor means of transportation. Nevertheless, it is hard to see how we could translate this evaluative judgment in terms of human agent normativity. Cars age and their components get worn-out; no designer has ever come up with an immortal car, so holding the car manufacturer responsible seems a bit forced. (S)he has done what (s)he had to do, at least within the boundaries of the current state of the art.

Now, these arguments might not to be decisive. I just have shown the problems one can encounter, when one denies artifacts any kind of normativity. Perhaps there are other arguments that deny artifactual normativity and avoid these issues. If so, I cannot find them. In the meantime, I think it is reasonable to take a modest, pragmatic position: our concept of artifactual norms has instrumental value, since it allows us to make better sense of engineering practice and engineering language. Therefore, the question whether they are genuinely normative is of minor importance and can be postponed until a definitive and complete account of normativity settles the issue.

3. What Normative Realists Ought to Reconsider: Part II

Normative realism doesn't make sense of artifactual normativity, I claimed. But, is its explanation of agent normativity satisfactory? I will contend it isn't. I will formulate two general points of critique. They may be not decisive, but will justify at least why I will develop an alternative (sections 4-6).

First, recall that *normative realism* adheres to the thesis that agents have good reasons to act in some ways rather than others in virtue of the existence of an *independent normative reality,* the latter consisting of reason giving options [11]. It claims that options *themselves* provide sufficient reasons merely *in virtue of their own nature*, irrespective of human make-up, desires,

[11] See Lillehammer 2002 and 2003.

rationality, and the like. In other words, the normative realist holds to a response-*in*dependent normative reality. To understand this view even better, let me contrast it with an account of response-*dependence*.

Response-dependent theorists hold to what we might call the *rational intelligibility condition*. This condition stipulates that options provide normative reasons only in virtue of being responded to by rational agents. Agents have normative reasons to pursue desires only on the condition that these desires would be endorsed in rationally favourable circumstances. Where the normative realist maintains that options provide normative reasons in virtue of their intrinsic nature — say, their intrinsic goodness — the response-dependency theorist will claim they do so in virtue of their external relations to the responses of agents to those options. Thus, while the goodness of an option is an intrinsic property for the normative realist, the response-dependent theorist defines it relative to the rationality of the agent confronted with it. In Michael Smith's words, an option is a good option in as far as a fully rational agent would desire it [12].

I will return to response-dependency in section 4. For now, I hope to have illustrated what it means to say that the normative realist holds to a response-*in*dependent normative reality.

Now, one problem of normative realism I will call the *autonomy* problem. It is related to the following [13]. Suppose normative realists are right and that normative reasons indeed are to guide human action. Then, on a purely response-*in*dependent account it would be hard to explain the fact that finite human beings are able to recognize them and to respond to them. As Lillehammer says [14]:

> If ends provide reasons in virtue of their nature, what is to stop this nature from being such as to outrun the best possible efforts of finite agents to grasp them as reason-giving in rational deliberation?

If humans are in no position to recognize reasons, then reasons lose their normative and practical function. It would be impossible for them to guide us, to improve, correct and evaluate our actions. In a sense, they are too *autonomous* to perform their supposed normative tasks.

[12] Smith, 2002, p. 329.

[13] Here I rephrase an argument of Lillehammer (2002, p.50).

[14] Lillehammer, 2003, p.4.

So what the *rational intelligibility condition* urges is that the extension of normative reasons be constrained by facts within the grasp of finite agents who reason soundly. Thus, the normative realist should make plausible that normativity involves at least some constraint(s) on the make-up of agents. Like Christine Korsgaard has argued: any realist account divorcing the existence of reasons from the exercise of a *capacity for practical rationality* fails to answer the normative question [15].

Lillehammer, however, doubts that the normative realist can do so [16]. Rather, he alleges that intermediate positions which claim to reconcile normative realism with a form of response-dependency are untenable. It is beyond the scope of this paper to scrutinize his arguments. Besides, I think normative realism faces yet another problem. It is what I call the problem of *intentionality*, and I think it speaks even more in favour of abandoning normative realism.

Defenders of normative realism often use phrasings such as: options or facts are normative reasons in as far they "exclude" [17], "favour" [18], "prescribe", and "contribute to" certain paths of action, insofar as they "speak" to us, insofar as they "tell" us what to do. I find these formulations pretty odd. If one says that a fact *favours* a certain way of going on*,* *how* does this favouring work? By what kind of magic does the fact that "playing with guns is wrong" *speaks* to us?

The point I want to make is the following: normative realists take a kind of Dennettian *intentional* stance towards facts and reasons, because they have to. If the clue to normativity is to be found in an independent normative reality, the latter actually has to *do* something; facts are to be interpreted as intentional agents who "speak" or "favour" or "prescribe." [19]

To be clear, I have no objections to intentional stances as such although I doubt that normative realists share my instrumentalist tendencies. In any case, I think that they should be more explicit about what they exactly hope

[15] See Korsgaard, 1995, p. 14ff.

[16] see Lillehammer, 2003.

[17] See my explanation of Raz's account, section 2.

[18] See my explanation of Dancy's account, section 2.

[19] And not only that, they should do it in such a way that finite human beings are able to receive the message. In speaking, the fact that "playing with guns is wrong" should keep, so to say, in the back of its mind, that its message should be recognizable to human beings. Maybe it's my lack of imagination, but I do not see how facts (be it stone-facts or so-called objective normative facts) can do all that.

to explain, when they intentionalize facts and reasons. Perhaps they are merely talking "metaphorically", but this metaphor should be explained since it has wide-ranging repercussions. For one thing, it puts their so-called realism in jeopardy and with it the existence of an independent normative reality, unless the latter consists of concepts instead of (normative) facts. Second, it is unclear what lies underneath their metaphorical talk. Consider Dancy's favouring relation: the normative fact that "playing with guns is dangerous" favours Jesse's not playing with guns. Is this favouring a causal relation? Dancy says no [20]. As a non-naturalist, he believes that the normative does not supervene on the descriptive; he thinks there is a normative realm, which does not necessarily correspond to a descriptive counterpart. Favouring is the basic normative relation and doesn't need further explanation; it just occurs. Why and how? We just don't know. (Nevertheless, *I* want to know.)

Both remarks concerning the problem of autonomy and intentionality lead me to the following conclusion: Normative realism is at best a theory to *make sense* of and *interpret* human action. It might indeed have some *instrumental* value and the concepts it offers — such as reasons, prescriptions, and the like — probably *are* used in our common-sense vocabulary and our folk morality. Nevertheless, I think it would be better to develop an interpretative strategy which has at least *some* underpinning in our natural world. It ought to be a theory which avoids what Blackburn calls a Platonic mystery, i.e., a theory which does not rely on '[normative] facts which bear only a strange relationship to the natural order, and whose own credentials and authority remain shrouded in obscurity [21]'. To such theory we turn now.

4. Oughts in Rational Explanations

Thus far I have told a negative story. To sum up, an alternative should meet three criteria:

> (A.i) it should explain artifactual normativity;
> (A.ii) it should avoid the autonomy problem by taking into account human dispositions;
> and
> (A.iii) if interpretative, intentional terms should somehow be backed up by a causal story.

[20] personal communication.

[21] Blackburn, 1998, p. 55.

I think a form of interpretative dispositionalism is a fair candidate. I will introduce the basic notions in this section, and apply them to technology in section 5.

In general, on a dispositional or response-dependent account of an entity, say a value, the nature and existence of that entity is constituted by the responses of agents to the world in some non-trivially defined set of favourable circumstances [22]. Consider again Michael Smith's dispositional theory of value: an option is a good option insofar as a fully rational agent would desire it. This means the goodness of an option is not an intrinsic property, but is defined by a human response [viz. a desire] in a set of favourable circumstances [viz. under the condition the agent is fully rational].

Now, if we want to *make sense of* an agent's behaviour, we can similarly refer to his distinctive psychological responses: roughly, by reference to the psychological states which reflect the information he has recorded and the inclination that moves him, in short, by reference to his beliefs and desires. Of course, such explanations in terms of mental states are not genuinely causal. Nevertheless, I take it to be uncontroversial that they are in some way related to a causal story, a story probably in neurophysiological terms – a form of causality I claimed to be lacking in normative realist accounts. In fact, we might say we have two kinds of explanations: a neurophysiological one, dwelling in the order of *causality*, and one in terms of the mental, dwelling in the order of *rationality* [23]. Thus, mental states are characterized by their place in a rational structure. And as Simon Blackburn says:

> [...] "rational" here means *normative*: it tells us how it would make sense for a person to factor a belief or desire into a pre-existent matrix of mental states. [...] It is a matter of 'rationalizing' the subjects, hypothesizing that they believe what they ought to believe, and desire what they ought to desire, or at least what it makes sense for them to desire [24].

This means we *interpret* human beings as creatures with beliefs, desires, and other states of mind who behave in ways that makes sense, given those states of mind. It must be clear that this interpretative strategy comes close to a form of Dennettian interpretationism. To be more specific, when

[22] See Lillehammer, 2003, p. 5.

[23] Both orders are close to what Daniel Dennett calls the physical and the intentional. See Dennett, 1987.

[24] Blackburn, 1998, p. 53-54, italics added.

approaching other intentional systems, we take a Dennettian Intentional Stance (IS):

> Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent [i.e. you put a rationality assumption in place, K.V.]; then you figure out what beliefs that agent *ought to have*, given its place in the world and its purposes. Then you figure out what desires it *ought to have*, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many — but not all — instances yield a decision about what the agent *ought to do*. [25]

Moreover, like Blackburn, Dennett discerns between causal and rational explanations. First, causal explanations are the outcome of taking the Physical Stance (PS). PS is an explanatory strategy which appeals to the physics of the explanandum — a particle, an object, an organism. Rational explanations, on the other hand, are the product of an IS. The latter is normative, since you explain, under the assumption of rationality, what a certain intentional system *ought to do* — as suggested by Dennett's quote above. Or, following Pettit [26], we can characterize it by a hypothetical imperative:

> (B) *if* an agent is to count as a rational being, given his beliefs and desires, he *ought* to act so-and-so.

First, note that this formulation is close to my suggestion (section 2) that norms are idealizations, putting comparative constraints on how things ought to be done or how they ought to be. Second, I think we already have met constraints (A.ii) and (A.iii). Recall that (A.ii) stated that we are in need of a theory which takes human dispositions into account, in order to avoid the so-called autonomy problem. The goodness of an action — the thing an agent ought to do — is, on our account, dependent on responses [viz. beliefs, desires] in a set of favourable circumstance [viz. under the condition is the agent is rational]. Moreover, (A.iii) is met as well. Our strategy is interpretative *and* the entities we pose have a causal counterpart. To understand this, we can refer again to Dennett. His PS and IS are *not* entirely unrelated. Intentional systems to be explained by taking an IS, are just a subset of all materially existing entities. And again, beliefs and desires as

[25] See Dennett, 1987, p. 17, italics added.
[26] See Pettit, 2002, p. 283.

used in IS, can to a certain extent be explained in physical terms, that is, by taken a PS towards them. Of course, I keep the PS/IS relationship rather vague, since a thorough analysis is, unfortunately, beyond the scope of this paper.

With this in mind, there remain two things to be done. First we need to make sense of (A.i) and we need to apply (A.i)-(A.iii) to technology. This is the subject of next section.

5. Norms in Technology: Explaining Human Agent and Artifactual Oughts

Let's start with human agent normativity. The hypothetical imperative (B) can be rephrased as follows: in explaining the actions of an agent, we put in place an *assumption* that, absent malfunction and other disturbing factors, the agent satisfies the role of a rational agent: he is more or less *rational* in its *responses* to evidence and more or less rational in moving from what he believes and from his values values to what to do [27]. Given an antecedent state, the agent *ought to do* X, on pain of being irrational; or, if he is to fulfill his role as a rational agent, he ought to respond according to the norms of rationality. With the aforementioned assumption in place, we are in a position to explain what the agent ought to do. Without it we would fail to explain, interpret or understand human behaviour.

Now, to explain the actions an *engineer* ought to perform, we first have to add that people satisfy the role of rational creatures as a result of natural selection *and* of cultural influence [28]. What it is for an engineer to be rational presumably differs from, say, scientific rationality, since both forms of rationality have evolved in different cultural niches. For instance, it might be rational for an engineer to act upon false beliefs, say as a heuristic to gain time; if a scientist would do the same, we usually would call him irrational. Rationality is a multi-faceted notion which co-varies with the conditions of its application, and this, in turn, has repercussions on our interpretations. It is reasonable to suppose that, when I interpret the actions of an engineer, I put a different rationality assumption in place, for instance, than when I interpret the actions of a scientist.

An obvious challenge to this approach is that I haven't said anything about what this engineering rationality consists in. I might seem to have shifted the

[27] See Cherniak, 1986 and Pettit, 2002.

[28] See Dennett, 1995, p. 506 and Pettit, 2002, p. 185.

normative question to a question of rationality. Nonetheless, I hope to have shed light on these issues. For one thing, I have suggested where to look for an answer and, as important, where *not* to look: in the realm of normative realism. Second, the purpose of the present paper was to overcome some problems of normative realism; it goes without saying that further research needs to flesh out the rationality assumption I use. In future work, I will offer an account of engineering rationality by altering Dennett's Stance Theory, which consists of three stances: the earlier mentioned PS and IS, and the Design Stance (DS). This last is an interpretative strategy to explain the behavior of designed entities, both biological and artificial[29]. I will introduce projective correlates to these stances. Engineers, I argue, do not *interpret* actual artifacts (as in DS) or actual users (as in IS), but try to *predict* the behavior of artifacts *which do not exist yet* and of *possible* users.

The remainder of this section, now, will concentrate on one problem I promised to solve: how to understand artifactual normativity. As for agent normativity, I will do this by means of interpretative strategies.

Let's start with an example [30]. Suppose we have designed a computer to add numbers presented to it and to display the sum: we have designed it to function as an adding device. The computer is a designed entity, so Dennett's DS applies. Thus, we ignore the details of the physical constitution of the object, assume that it has been designed with a certain purpose in mind, and explain its behavior accordingly. Instead of working under the assumption of rationality (as in IS) however, we put an optimality assumption[31] in place: the artifact was designed so that it actually can perform its intended function. So, if our design is successful, whenever we present the computer with a set of numbers, it will respond by giving us their sum.

As in the case of human action, the sort of regularity involved in the computer's responses has the status of a norm. To understand this, we can invoke a hypothetical imperative again (cfr. (B)): if the machine we developed is to count as an adder, for input seven and four, it *ought* to produce output eleven. Or, under the *assumption* that the system is an adder, we can say that it ought to output the sum of the inputs, where the ought is a *normative* ought. In this case the norm refers to the ideal state in which the

[29] For Dennett biological objects indeed can be considered as designed entities. Interpreting them means we take a kind of intentional stance towards 'Mother Nature', as he calls the process of natural selection.

[30] The example is taken from Pettit, 2002.

[31] In our example such optimality assumption is in fact an effectiveness assumption. For sake of clarity we however stick to Dennett's original terminology.

artifact functions properly. In light of this optimality standard, we interpret and evaluate its functioning.

Artifactual normativity, however, is not restricted to the artifact's capacity to perform its intended function. Take, for instance, the statement: "a car ought to be safe and clean." We can explain its normativity, again by invoking a hypothetical imperative: for a car to be marketable, it *ought to be* so-and-so; it ought to be safe, it ought to be clean, it ought to comply to certain technical standards and the like.

To sum up, I hope to have shown with sections 4 and 5, that, contra normative realism, it makes sense to talk about artifactual normativity; moreover, I think my account is better suited than its normative realist counterpart, when it concerns the explanation of "engineer normativity".

Before concluding (section 7), I will take up some ontological issues with respect to my proposal.

6. Ontology And Objectivity Are Not Endangered Species

This section sketches briefly my ontological commitments. In particular it concerns two questions: (i) does my approach to normativity refer to anything at all?; and (ii) does it exclude the objectivity of norms?

The answer to the first question is: I hope so, but my account doesn't stand or fall with it. The approach I have defended is in both cases [i.e. human agent and artifactual normativity] explanatory in nature. It offers an interpretative strategy, without much of an ontological commitment — maybe apart from the fact that our explanations almost certainly depend on lower-level, causal explanations. In any case, we are interested primarily in the instrumental value of normative theories to the neglect of realist concerns and we believe that our account helps make sense of engineering normativity. I won't illustrate this contention for the case of "human agent normativity" — I think, for instance, Dennett has sufficiently done so — but focus on artifactual normativity instead.

One benefit of my approach is the following. We can learn that a certain system is designed or selected to fit a certain role and we can determine its normative regularities, without having to know the regularities of its lower-level causal structure. Knowing the designer or his purposes, or just a little empirical evidence of the system itself, may convince us that this system, say, is a device meant to add. And with this in mind, we are in a position to predict its behavior, absent malfunctioning. Second, such explanations have evaluative value. *If* a certain system is to count as an adder, it ought to be designed in a way that it gives the correct sum when presented with a set of numbers. If it doesn't fulfill this role, it is a bad adder or a malfunctioning one. Third, noticing some regularities in a system might direct us in finding answers to the causal story or history which has brought them about. We might analyze whether and how these regularities were programmed for. Stated differently, the higher-level interpretation of a system's behavior may be of guidance to study its lower-level counterparts.

Now, turning to the second question, I do think my account holds to a certain objectivity of norms. In the case of human agents we work under rationality assumptions, in the case of artifacts we invoke optimality assumptions. Both "rationality" and "optimality" are crucial to our normative claims, that is, they constrain what we reasonably can expect persons and things to do or to be like. For instance, suppose I interpret a person. What he ought to do is not merely dependent on his *individual* beliefs and desires, but also on what he *as a rational being*, given his mental states, is supposed to do. As such, the notion of rationality can be used to define better or worse ways of responding to a certain situation. Not any response will do. Whether this is sufficient to be called "genuine" objectivity, I do not know. At least it is *not* the objectivity, normative realists are after: a set of platonic norms mysteriously trying to persuade us to do this-and-that.

1. Conclusion

To give more structure to the story I have told, I will sum up its main contentions. With this paper I hope to have demonstrated that:

1. Normative realism offers at best an interpretative strategy to understand normativity. On one hand this seems incompatible with its realist ambitions. On the other hand, as an interpretative strategy it falls short in two respects: (i) as presented by Raz and Dancy it fails to account for artifactual normativity; and (ii) it lacks a supporting causal story for human agent normativity.
2. If one is to explain the normativity in technology, one has to embrace a kind of dispositional interpretationism. What an engineer ought to do is explained in terms of his responses [viz. his desires and beliefs] under the assumption that he is rational; he ought to do X, on pain of being irrational. What an artifact ought to do is explained in terms of its responses under the assumption that it purports to fulfill its artifactual role; it ought to perform Y, on pain of failing to be a well-functioning artifact of type Z. A similar strategy applies when we

interpret the non-functional normative constraints on artifacts, for instance when we make claims about how an artifact ought to be like.

3. The challenge to this account is to analyze more thoroughly the rationality assumption in the case of engineering actions. Nevertheless, I briefly argued how we could proceed from here on.

References

Blackburn, Simon. 1998. Ruling Passions: a Theory of Practical Reasoning. Oxford Clarendon Press.

Castañeda, Hector-Neri. 1970. On the Semantics of the Ought-to-Do. *Synthese* 21: 449-468.

Cherniak, Christopher. 1986. Minimal Rationality. MIT Press.

Dancy, Jonathan. 2004. Ethics Without Principles. Oxford University Press.

Dennett, Daniel Clement. 1987. The Intentional Stance. MIT Press.

———. 1995. Darwin's Dangerous Idea: Evolution and the Meanings of Life. Simon & Schuster.

Korsgaard, Christine. 1995. The Sources of Normativity. Cambridge University Press.

Lillehammer, Halvard. 2002. Moral realism, Normative Reasons, and Rational Intelligibility. *Erkenntnis* 57: 47-69.

———. 2003. The Metaphysics of Normative Reasons. in *Grundlagen Der Ethik: Normativität und Objektivität*, edited by ed. Schaber, P. & R. Hüntelman. Frankfurt: Ontos Verlag.

Pettit, Philip. 1996. The Common Mind: An Essay on Psychology, Society and Politics. Oxford University Press.

———. 2002. Three Aspects of Rational Explanation. In *Rules, Reasons and Norms*. Oxford University Press.

Smith, Michael. 2002. Exploring the Implications of the Dispositional Theory of Value. *Philosophical Issues*, 12, p. 329-347.

Raz, Joseph. 1975 & 1990. Practical Reason and Norms. Princeton University Press.

———. 1999. Engaging Reason: On the Theory of Value and Action. Oxford University Press.