

PROCEEDINGS IACAP 2011

---

FIRST INTERNATIONAL CONFERENCE OF  
IACAP

# THE COMPUTATIONAL TURN: PAST, PRESENTS, FUTURES?

4 – 6 JULY, 2011

AARHUS UNIVERSITY

**PRINTED WITH THE FINANCIAL SUPPORT OF THE HEINZ NIXDORF  
INSTITUTE, UNIVERSITY PADERBORN, GERMANY**

**© VERLAGSHAUS MONSENSTEIN UND VANNERDAT OHG  
AM HAWERKAMP 31  
48155 MÜNSTER**

**“The Computational Turn: Past, Presents, Futures?”**

Dear participants,

In the West, philosophical attention to computation and computational devices is at least as old as Leibniz. But since the early 1940s, electronic computers have evolved from a few machines filling several rooms to widely diffused – indeed, ubiquitous – devices, ranging from networked desktops, laptops, smartphones and “the internet of things.” Along the way, initial philosophical attention – in particular, to the ethical and social implications of these devices (so Norbert Wiener, 1950) – became sufficiently broad and influential as to justify the phrase “the computational turn” by the 1980s. In part, the computational turn referred to the multiple ways in which the increasing availability and usability of computers allowed philosophers to explore a range of traditional philosophical interests – e.g., in logic, artificial intelligence, philosophical mathematics, ethics, political philosophy, epistemology, ontology, to name a few – in new ways, often shedding significant new light on traditional issues and arguments. Simultaneously, computer scientists, mathematicians, and others whose work focused on computation and computational devices often found their work to evoke (if not force) reflection and debate precisely on the philosophical assumptions and potential implications of their research. These two large streams of development - especially as calling for necessary interdisciplinary dialogues that crossed what were otherwise often hard disciplinary boundaries – inspired what became the first of the Computing and Philosophy (CAP) conferences in 1986 (devoted to Computer-Assisted Instruction in philosophy).

Since 1986, CAP conferences have grown in scope and range, to include an extensive array of intersections between computation and philosophy as explored across a global range of cultures and traditions – issuing in fruitful cross-disciplinary collaborations and numerous watershed insights and contributions to scholarly reflection and publication. In keeping with what has now become a significant tradition of critical inquiry and reflection in these domains, IACAP'11 celebrates the 25th anniversary of CAP conferences by focusing on the past, present(s), and possible future(s) of the computational turn.

Aarhus, July 2011

Charles Ess  
Organizer  
Department of Information- and Media Studies  
Aarhus University

Ruth Hagenruber  
Program Chair  
Paderborn University

## ACKNOWLEDGEMENTS

Happily, in planning and organizing IACAP'11, I have received generous support and encouragement from more persons and institutions than can be fully listed here – beginning with the Track Chairs, members of the Program Committee / Comité scientifique, and the keynote speakers who have kindly accepted our invitation to join us in Aarhus for our conference.

In addition, I would like to express deep gratitude to my colleagues in the Department of Information- and Media Studies (IMV), Aarhus University, including the highly competent members of the secretariat and our chair, Steffen Ejnar Brandorff. Without your on-going encouragement, assistance, and financial support, IACAP'11 would simply not have taken place at Aarhus University.

I am also very grateful to Aarhus University for additional forms of support, including their conference facilities and most especially the very able assistance of Ulla Rasmussen Billings (Faculty Secretariat) for her assistance and advice on multiple conference matters, including budgeting and the conference registration page.

For the first time in its now 25-year history, IACAP has offered travel bursaries to support the participation of our younger colleagues: Dr. Johnny Søraker has ably taken on the difficult chore of coordinating the awarding of these bursaries. Many thanks (*mange tak!*).

Finally, a thousand thanks (*tusind tak!*) to Prof. Dr. Ruth Hagengruber (Universität Paderborn) who has undertaken not only the daunting role of Program Chair, but also for editing, and producing these Proceedings for IACAP'11.

Aarhus, July 2011

Charles Ess

## Table of Contents

### Keynotes

#### *Presidential address*

Beavers, Anthony F. 19  
IS ETHICS COMPUTABLE, OR WHAT  
OTHER THAN *CAN* DOES *OUGHT* IMPLY

Aas, Katja Franko 21  
(IN)SECURE IDENTITIES: ICTS, TRUST  
AND BIOPOLITICAL TATTOOS

#### *Covey Lifetime Achievement Award*

Bynum, Terrell Ward 22  
INFORMATION AND DEEP  
METAPHYSICS

#### *Herbert A. Simon Award for Outstanding Research in Computing and Philosophy*

Sullins, John P. 24  
THE NEXT STEPS IN ROBOETHICS

#### *Brian Michael Goldberg Award for Outstanding Graduate Research in Computing and Philosophy (sponsored by Carnegie Mellon University)*

Buckner, Cameron 25  
COMPUTATIONAL METHODS FOR THE  
21ST CENTURY PHILOSOPHER: RECENT  
ADVANCES AND CHALLENGES IN  
COGNITIVE SCIENCE AND  
METAPHILOSOPHY

**Panel**

Charles Ess / Elizabeth Buchanan / Jeremy Mauger	INTERNET RESEARCH ETHICS INTERNET RESEARCH ETHICS: CORE CHALLENGES, NEW DIRECTIONS	26
--	--	----

**Tracks**

**Track I: Philosophy of Computer Science**

Bengez, Rainhard Z.	RULES AND PROGRAMMING LANGUAGES	29
Blanco, Javier O. et alia	A BEHAVIOURAL CHARACTERIZATION OF COMPUTATIONAL SYSTEMS	30
Boltuc, Peter	WHAT IS THE DIFFERENCE BETWEEN YOUR FRIEND AND A CHURCH TURING LOVER	34
Chokvasin, Theptawee	HAECITY AND INFORMATION	37
Duran, Juan M.	THE LIMITS OF COMPUTER SIMULATIONS AS EPISTEMIC TOOLS	40
Franchette, Florent	WHY TO BUILD A PHYSICAL MODEL OF HYPERCOMPUTATION	43
Geier, Fabian	THE MATERIALISTIC FALLACY	46

Meyer, Steven	THE EFFECT OF COMPUTERS ON UNDERSTANDING TRUTH	49
Monin, Alexandre, Halpin, Harry	PHILOSOPHY OF THE WEB AS ARTIFACTUALIZATION	53
Pagano, Miguel	ONTOLOGICAL COMMITMENTS OF COMPUTER SCIENCE	54
Riss, Uwe	SEMANTICS OF PROGRAMMING LANGUAGES	60
Sinclair, Nathan	QUINEAN HOLISM AND THE INDETERMINANCY OF COMPILATION	64
Smith, Lindsay	IS FINDING A ‚BLACK SWAN‘ POPPER, (1936) POSSIBLE IN SOFTWARE DEVELOPMENT?	67
Solodovnik, Iryna	ONTOLOGY: FROM PHILOSOPHY TO ICT AND RELATED AREAS. PROBLEMS AND PERSPECTIVES	71
Thürmel, Sabine	THE EVOLUTION OF SOFTWARE AGENTS AS DIGITAL OBJECTS	74
Turner, Raymond	MACHINES AND COMPUTATIONS	77
 <b>Track II: Philosophy of Information and Cognition</b>		
Funcke, Alexander	ON THE LEVEL OF CREATIVITY. PONDERINGS ON THE NATURE OF KANTIAN CATEGORIES, CREATIVITY AND COPYRIGHTS	79
Giardino, Valeria	THE FOURTH REVOLUTION AND SEMANTIC INFORMATION	83

Heersmink, Richard	EPISTEMOLOGICAL AND PHENOMENOLOGICAL ISSUES IN THE USE OF BRAIN-COMPUTER INTERFACES	87
Hewlett, David, Cohen, Paul	AN INFORMATION-THEORETIC MODEL OF CHUNKING	91
Janlert, Lars-Erik	THE DYNAMISM OF INFORMATION ACCESS FOR A MOBILE AGENT IN A DYNAMIC SETTING AND SOME OF ITS IMPLICATIONS	94
Kitto, Kirsty	CONTEXTUAL INFORMATION: MODELING DIFFERENT INTERPRETATIONS OF THE SAME DATA WITHIN A GEOMETRIC FRAMEWORK	97
Menant, Christophe	COGNITION AS A MANAGEMENT OF MEANINGFUL INFORMATION: PROPOSAL FOR AN EVOLUTIONARY APPROACH	101
Quiroz, Francisco Hernandez	COMPUTATIONAL AND HUMAN MIND MODEL	104
Schroeder, Marcin	SEMANTICS OF INFORMATION: MEANING AND TRUTH AS RELATIONSHIPS BETWEEN INFORMATION CARRIERS	107
Vakarelov, Orlin	PRE-COGNITIVE SEMANTIC INFORMATION	111

**Track III: Autonomous Robots and Artificial Cognitive systems**

Anokhina, Margaryta, Dodig- Crnkovic, Gordana	WHO WILL HAVE IRRESPONSIBLE, UNTRUSTWORTHY, IMMORAL INTELLIGENT ROBOT? WHY ARTIFACTUALLY INTELLIGENT ADAPTIVE AUTONOMOUS AGENTS NEED TO BE ARTIFACTUALLY MORAL?	115
Arkin, Ronald	THE ETHICS OF ROBOTIC DECEPTION	118
Bello, Paul et alia	PROLEGOMENON TO ANY FUTURE THEORY OF MACHINE AUTONOMY	121
Briggs, Gordon	AUTONOMOUS AGENTS AND SENSES OF RESPONSIBILITY	124
Hagengruber, Ruth	THE ENGINEERABILITY OF SOCIAL INSTITUTIONS	127
Heimo, Olli I., Kimppa, Kai K.	RESPONSIBILITY IN ACQUIRING CRITICAL EGOVERNMENT SYSTEMS: WHOSE FAULT IS FAILURE?	129
Kavathatzopoulos, Iordanis, Laaksoharju, Mikael Molyneux, Bernard	WHAT ARE ETHICAL AGENTS AND HOW CAN WE MAKE THEM WORK PROPERLY?	133
Vallverdu, Jordi, Casacuberta, David	HOW THE HARD PROBLEM OF CONSCIOUSNESS MIGHT ARISE FOR AN EMBODIED (SYMBOL) SYSTEM	136
Veale, Richard	THE GAME OF EMOTIONS (GOE): AN EVOLUTIONARY APPROACH TO AI DECISIONS	139
Waser, Mark R.	THE CASE FOR DEVELOPMENTAL NEUROBOTICS	143
	WISDOM DOES IMPLY BENEVOLENCE	148

**Track IV: Technosecurity from Every day Surveillance to Digital Warfare**

Crutzen, C.K.M.	THE MASKING AND UNMASKING OF PRIVACY	152
Hempel, Leon	CHANGE AND CONTINUITY – FROM THE CLOSED WORLD OF BIPOLARITY TO THE CLOSED WORLD OF THE PRESENT	155
Macnish, Kevin	SUBITO AND THE ETHICS OF AUTOMATING THREAT ASSESSMENT	159
Othmer, Julius, Weich, Andreas	MATCHING – POPULAR MEDIA BETWEEN SECURITYWORLDS AND CULTURES OF RISK	162
Taddeo, Mariarosa	INFORMATIONAL WARFARE AND JUST WAR THEORY	164
Weber, Jutta	TECHNO-SECURITY, RISK AND THE MILITARIZATION OF EVERY DAY LIFE	168

**Track V: Information Ethics, Robot Ethics**

Asaro, Peter	IS THERE A HUMAN RIGHT NOT TO BE KILLED BY A MACHINE?	175
Dasch, Thomas	DO WE NEED AN UNIVERSAL INFORMATION ETHICS?	177
Douglas, Keith	A PSEUDOPERIPATETIC APPLICATION SECURITY HANDBOOK FOR VIRTUOUS SOFTWARE	180

Hromada, Daniel D.	THE CENTRAL PROBLEM OF ROBOETHICS: FROM DEFINITION TOWARDS SOLUTION	182
Soraker, Johnny Hartz	AFFECTING THE WORLD OR AFFECTING THE MIND? THE ROLE OF MIND IN COMPUTER ETHICS	186
Tonkens, Ryan	THE ETHICS OF AUTOMATED WARFARE	190
Vallor, Shannon	CAREBOTS AND CAREGIVERS: ROBOTICS AND THE ETHICAL IDEA OF CARE	193
Wong, Pak-Hang	CO-CONSTRUCTION AND CO- MANAGEMENT OF ONLINE IDENTITIES: A CONFUCIAN PERSPECTIVE	197
 <b>Track VI: Multidisciplinary Perspectives</b>		
Baumgaertner, Bert	REFLECTIVE INEQUILIBRIUM	202
Belfer, Israel	THE INFORMATION-COMPUTATION TURN: A HACKING-TYPE REVOLUTION	205
Breems, Nick	COMPUTERS AND PROCRASTINATION: „I’LL JUST CHECK MY FACEBOOK QUICK A SECOND“	209
Bod, Rens et alia	HOW MUCH DO FORMAL NARRATIVE ANNOTATIONS DIFFER? A PROPRIAN CASE STUDY	212
Desclés, Jean- Pierre et alia	COMBINATORY LOGIC WITH FUNCTIONAL TYPES IS A GENERAL FORMALISM FOR COMPUTING COGNITIVE AND SEMANTIC REPRESENTATIONS	216

Franchi, Stefano	THE PAST, PRESENT AND FUTURE ENCOUNTERS BETWEEN COMPUTATIONS AND THE HUMANITIES	219
Guarini, Marcello et alia	REFLECTIONS ON NEUROCOMPUTATIONAL RELIABILISM	224
McKinley, Steve	STATES OF AFFAIRS AND INFORMATION OBJECTS	227
McKinley, Steve	SCIENTIFIC EXPLANATION AND INFORMATION	230
Nicolaidis, Michael	BIOLOGICAL INSPIRED SINGLE-CHIP MASSIVELY PARALLEL SELF-HEALING,  SELF-REGULATING, TERA-DEVICE COMPUTERS: PHILOSOPHICAL IMPLICATIONS OF THE EFFORTS FOR SOLVING TECHNOLOGICAL  SHOW-STOPPERS IN THE PATH OF THE NEXT COMPUTATIONAL TURN	234
Portier, Pierre- Edouard, Calabretto, Sylvie	STRUCTURAL CONSTRAINTS FOR THE CONSTRUCTION OF MULTI- STRUCTURED DOCUMENTS	238
York, William W., Ekbia, Hamid R.	(DIS)TASTEFUL MACHINES? AESTHETIC COGNITION AND THE COMPUTATIONAL TURN IN AESTHETICS	243

**Track VII: Social Computing**

Alhutter, Doris	THE SOCIAL AND ITS POLITICAL DIMENSION IN SOFTWARE DESIGN: A SOCIO-POLITICAL APPROACH	248
Barker, Steve	A SOCIAL EPISTEMOLOGICAL APPROACH FOR DISTRIBUTED COMPUTER SECURITY	251
Coeckelbergh, Mark	TRUST, POWER AND INFORMATION TECHNOLOGY	254
Compagna, Diego	THE BENEFITS OF SOCIAL THEORY FOR MODELLING STABLE ENVIRONMENTS OF SYSTEMIC TRUST WITHIN MULTI AGENT SYSTEMS	258
Danka, Istvan	COMPUTER NETWORKS AND THE PHILOSOPHY OF MIND. A SOCIAL MIND – NETWORKED COMPUTER ANALOGY	260
Dodig-Crnkovic, Gordana	AGENT BASED MODELING WITH APPLICATIONS TO SOCIAL COMPUTING	262
Ekbia, Hamid R., Zhang, Guo	OBJECTS OF IDENTITY, IDENTITY OF OBJECTS: FOR A MATERIALIST ACCOUNT OF ONLINE BEHAVIOUR	265
Ropolyi, Laszlo	THE CONSTRUCTION OF REALITY AND OF SOCIAL BEING IN THE INFORMATION AGE	269
Simon, Judith	TRUST, KNOWLEDGE AND SOCIAL COMPUTING. RELATING PHILOSOPHY OF COMPUTING AND EPISTEMOLOGY	272
Vehlken, Sebastian	OPERATIONAL IMAGES. AGENT-BASED COMPUTER SIMULATIONS AND THE EPISTEMIC IMPACT OF DYNAMIC VISUALIZATION	275

Zambak, Aziz	SOCIAL COMPUTATION AS A DISCOVERY MODEL FOR THE SOCIAL SCIENCES	278
--------------	---	-----

**Track VIII: IT, Culture and Globalization**

Asai, Ryoko et alia	THE REVIVAL OF NATIONAL AND CULTURAL IDENTITY THROUGH SOCIAL MEDIA	283
Backhaus, Patrick, Dodig-Crnkovic, Gordana	WIKILEAKS AND ETHICS OF WHISTLE BLOWING	286
De Gooijer, Thijmen	INTERPRETING CODES OF ETHICS IN GLOBAL SOFTWARE ENGINEERING	289
Hongladarom, Sonja	INFORMATION, TECHNOLOGY, GLOBALIZATION AND INTELLECTUAL PROPERTY RIGHTS	294

**Track IX: Surveillance, sousveillance...**

Beinsteiner, Andreas	TOWARDS A HERMENEUTIC PHENOMENOLOGY OF CYBER-SPACE: POWER VS. CONTROL	297
Ganascia, Jean- Gabriel	THE WIKILEAKS LOGIC	300
Najar, Anis	DEMOCRACY 2.0 – HOW THE WEB MAKES REVOLUTION	303
Reynolds, Carson	NEGATIVE SOUSVEILLANCE	306

Strauss, Stefan	GOVERNMENT APPROACHES FOR MANAGING ELECTRONIC IDENTITIES OF CITIZENS – EVOKING A CONTROL DILEMMA?	309
<b>Track X: SIG Track – Machines and Mentality</b>		
Arkin, Ronald C.	MORAL EMOTIONS FOR ROBOTS	313
Arkoudas, Konstantine	ON DEEPLY UNCONSCIOUS INTENTIONAL STATES	316
Bridewell, Will et alia	OUTLINING A COMPUTATIONALLY PLAUSIBLE APPROACH TO MENTAL STATE ASCRIPTION	319
Guarini, Marcello	AGENCY: ON MACHINES THAT MENTALIZE	322
Nirenburg, Sergej	TOWARD A TESTBED FOR MODELING THE KNOWLEDGE, GOALS AND MENTAL STATES OF OTHERS	325
Scheutz, Matthias	ARCHITECTURAL STEPS TOWARDS SELF-AWARE ROBOTS	328
Sundar, Naveen, Bringsjord, Selmer	LOGIC-BASED SIMULATIONS OF MIRROR TESTING FOR SELF- CONSCIOUSNESS	331
<b>List of Authors in Alphabetic Order</b>		334



# Keynotes

**IS ETHICS COMPUTABLE, OR WHAT OTHER THAN  
CAN DOES *OUGHT* IMPLY?**

**ANTHONY F. BEAVERS**  
*Department of Philosophy*  
*The University of Evansville*

In 2007, Anderson and Anderson wrote, “As Daniel Dennett (2006) recently stated, AI ‘makes philosophy honest.’ Ethics must be made computable in order to make it clear exactly how agents ought to behave in ethical dilemmas” (16). To rephrase, a computable system or theory of ethics makes ethics honest. But at what cost? Might Turing’s 1950 prophecy that “at the end of the century the use of words ... will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted” (1950, 442) soon take on normative dimensions due to research in artificial morality. Will attempts to make ethics computable lead us to redefine the term “moral” to fit the case of machines and thus change its meaning for humans also? I call this the threat of “moral nihilism ... the doctrine that states that morality needs no internal sanctions, that ethics can get by without moral “weight,” i.e., without some type of psychological force that restrains the satisfaction of our desire and that makes us care about our moral condition in the first place” (Beavers, 2011a).

Analyzing this possibility requires inspection of the meaning of the term “ought” and what it implies. In 2009, I argued that, following Kant, *ought* not only implies *can*, but also *might not*, in which case it would be morally wrong to create artificial Kantian agents, since doing so would require designing them in such a way that they *could* act immorally, but would not do so. Only on such a condition would it make sense to hold a machine responsible for its actions and praise or blame it for its behavior. In 2011, I argued that if *ought* implies *can*, then it also implies *implementability*. If a machine or human *can* act morally, this can only be because the mechanisms (whether in software or wetware) have the requisite components to allow for it. Thus, any theory of morality must be implementable in real working agents to qualify as a viable moral theory. Given the conclusions of 2009, I argued in 2011 that designing machines in such a way that they behaved morally but were not able to act immorally would require redefining the term “morality” in such a way that full moral agency with internal sanctions was not intrinsic to ethics, but “merely a

sufficient, and no longer necessary, condition for being ethical.” In this case, internal states such as conscience, responsibility (as felt affective weight) and thus moral accountability are, *ex hypothesi*, not necessary for ethics either. Thus, if we build machines capable of being described by the term “moral” we can only do so by redefining the term. So, if a time is coming when we can speak of a machine as moral without expecting to be contradicted, we will have succeeded in turning ethics into a strictly extrinsic, behavioral affair in which internals are irrelevant.

Since on the surface, an ethics without an *ought* is as empty as *thinking* without *insight* or *wisdom*, it is necessary to explore what else *ought* implies in order to form an adequate conception of a metaphysics of morals that will fit the information age. While other research for a working conception of ethics has already been done (e.g., Floridi and Sanders, 2004), a careful exploration of this foundational concept still appears lacking. I hope to fill this gap to explore whether ethics can get by without its cherished *ought* and, if so, what that implies for ethics more generally. The concern guiding this talk is whether the information age is issuing in a post-ethical age or whether it is leading to a redefinition of ethics that is both long overdue and needed.

## References

- Anderson, M., & Anderson, S. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4): 15-26.
- Beavers, A. (2011). Moral machines and the threat of ethical nihilism. In P. Lin, G. Bekey & K. Abney (Eds.), *Robot ethics: The ethical and social implication of robotics*. Cambridge, MA: MIT Press, forthcoming.
- Beavers, A. (2009, March). Between angels and animals: The question of robot ethics, or is Kantian moral agency desirable. The Eighteenth Annual Meeting of the Association for Practical and Professional Ethics, Cincinnati, Ohio.
- Dennett, D. (2006, May). Computers as prostheses for the imagination. The International Computers and Philosophy Conference, Laval, France.
- Floridi, L., & Sanders, J. (2004). On the morality of artificial agents. *Minds and Machines* 14(3): 349-379.
- Turing, A. (1950). Computing machinery and intelligence. *Mind* 59: 433-460.

**(IN)SECURE IDENTITIES: ICTS, TRUST AND ‘BIO-POLITICAL’  
TATTOOS**

**KATJA AAS**

*Department of Criminology and Sociology of Law  
University of Oslo*

The globalising world is marked by anonymity, mass mobility and mass consumerism. These conditions create a distinct set of challenges for social identification practices, first and foremost, the challenge of creating reliable and ‘trustworthy’ identities. The paper addresses in particular the growing reliance on biometrics and biometric databases and examines how these forms of bodily control function as border controls. While revealing specific notions of subjectivity, the paper also explores how these technologies function as mechanisms of social sorting and global governance and have markedly different effects on the citizen of the global North and the global South.

## INFORMATION AND DEEP METAPHYSICS

**TERRELL WARD BYNUM**

*Department of Philosophy*

*Southern Connecticut State University*

Scientists working on the cutting edges of their field often engage in thinking that is much like metaphysics. Similarly, in the past, philosophers inspired by major advances in science have made significant additions to metaphysics, as well as other branches of philosophy. On occasion, the scientists and philosophers have been the very same people. For example in ancient times Aristotle created physics, biology and animal psychology, and at the same time he made related contributions to metaphysics, logic, epistemology, and other branches of philosophy. Again, during the Enlightenment in Europe, influential philosophers like Descartes and Leibniz also were respected scientists and first-class mathematicians. At times, people who were primarily scientists (for example, Copernicus, Galileo, and Newton) inspired thinkers who were primarily philosophers (for example, Hobbes, Locke, and Kant). In more recent times, revolutionary scientific contributions of Darwin, Einstein, Schrödinger, Heisenberg, and others significantly influenced philosophical ideas of people like Spencer, Russell, Whitehead, Popper, and many more.

Today, in the early years of the twenty-first century, developments in cosmology and quantum physics appear likely to alter significantly our scientific understanding of the universe, of life, and of the human mind; and many scientists have become convinced that the universe, ultimately, is made of quantum information. These developments, it seems to me, are very likely to lead to important new contributions to philosophy; and indeed, as illustrated by Luciano Floridi's writings on informational realism and philosophy of information, significant philosophical contributions already have begun to appear.

Of special interest, in this presentation is the idea that *the universe is a vast "ocean" of quantum bits ("qubits")*; and thus each object or process in the universe can be seen as *a constantly changing data structure comprised of qubits*. On this account of the ultimate nature of the universe, the fundamental "stuff" of which our universe is made is *quantum information*. Unlike traditional "bits", such as those processed in most of today's information technology devices, "qubits" have quantum features such as genuine randomness, superposition and entanglement – features that Einstein and other

scientists considered “spooky” or “weird”. These nontraditional features of qubits have made it possible to achieve unbreakable encryption, teleportation, and a new kind of computing – “quantum computing”.

In this presentation, a number of quantum topics, such as randomness, superposition, entanglement, collapse of a wave function, teleportation, and quantum computing are briefly described. In light of such quantum features, it seems appropriate for philosophers to re-examine a variety of philosophical concepts, such as possibility and impossibility, potential and actual, cause and effect, being and reality, logic and contradiction, and a number of others. Such concepts are central to the “deep metaphysics” that provides a conceptual foundation for philosophy. Consequently, this presentation calls upon philosophers to familiarize themselves with current developments in cosmology and quantum physics, especially those developments that see the universe as ultimately an expanding ocean of quantum information. If philosophers take on this challenge – as Luciano Floridi has already begun to do – the deep metaphysical foundations of philosophy are likely to be profoundly transformed. As a small contribution to that effort, this presentation concludes with a brief sketch of a possible new metaphysical theory.

## **THE NEXT STEPS IN ROBOETHICS**

**JOHN P. SULLINS**

*Department of Philosophy  
Sonoma State University*

RoboEthics has now matured from its beginnings as a curious offshoot of computer ethics into a sub-discipline of its own that has a well defined scope of study. In this paper I will briefly look at the growth of RoboEthics and the important roll it is playing in the development of robotics technology. I will then look at the more pressing open problems in RoboEthics and suggest some ways forward. I will focus primary on the criticism that RoboEthics is impossible given that phronesis is beyond the capacity of machines. To refute this claim I will propose a model system inspired by the architecture of the IBM Watson computer that, I will argue, could achieve an artificial practical wisdom. This would be possible through the use of a context sensitive hybrid of logical and non-logical search methods that could access documents to find comparable exemplar cases similar to the ethical situation the robot is attempting to reason about. Armed with this data, the robot would be able to make more nuanced decisions even without its own innate human equivalent practical wisdom.

**COMPUTATIONAL METHODS FOR THE 21ST-CENTURY  
PHILOSOPHER: RECENT ADVANCES AND CHALLENGES IN  
COGNITIVE SCIENCE AND METAPHILOSOPHY**

**CAMERON BUCKNER**

*Department of Philosophy  
Indiana University*

As evidenced by past CAP conferences, the intersection of computing and philosophy has long been a fertile area of research. The past ten years in particular have produced a variety of new computational techniques of philosophical import. These powerful new techniques present philosophers with alluring opportunities, but also pose a number of challenges requiring methodological reforms. In cognitive science, new computational models of psychological processes are rapidly-increasing our ability to predict behaviors, but the structure of these models seem to make a hash of traditional distinctions in psychology such as that between cognition and association. In metaphilosophy, new statistical and logical programming methods offer the possibility to address otherwise intractable philosophical questions, but rely upon a variety of assumptions, require input data that can be expensive to collect, and produce results that can be difficult to evaluate. In this talk, I will review some of these new technologies, recommending new conceptual frameworks and methodologies to understand, evaluate, and utilize their results. While I will give a brief overview of this latest generation of research, the talk will focus primarily on specific examples from my own work in the areas of comparative psychology and dynamic ontology.

# Panel

## **INTERNET RESEARCH ETHICS: CORE CHALLENGES, NEW DIRECTIONS**

### **Charles Ess**

*Department of Information- and Media Studies  
Aarhus University*

### **Elizabeth Buchanan**

*Director, Center for Applied Ethics  
University of Wisconsin-Stout  
Co-Director, International Society for Ethics & Information  
Technology (INSEIT)*

### **Jeremy Mauger**

*School of Information Studies  
University of Wisconsin, Milwaukee*

Internet Research Ethics (IRE) is an emerging cross-disciplinary field which studies how research is conducted in online environments and seeks to resolve the subsequent ethical dilemmas in normative and practical terms. While similar to its physical counterpart, conducting scholarly research online is different in terms of ethics and values. For example, online surveys bring new privacy concerns. Research in chat rooms confounds our notions of subject anonymity and identifiability. Scraping data from social networks or public blogs complicates issues of informed consent. At the same time, research conducted on and through the Internet has expanded exponentially in the last ten years; researchers across disciplines make frequent use of such tools as online survey generators, as well as engage in forms of participant observations of virtual worlds. Internet Research Ethics has thus emerged over the past decade as a distinct and important field of applied ethics – one that overlaps with central issues and approaches of information and computing ethics and is often informed (and informs) the broader intersections between computing and philosophy.

The panel will begin with a few real-world examples of ethical dilemmas that are representative of contemporary issues in IRE and are especially challenging to traditional ethics. Panelists will then provide an overview of two current projects focusing on significantly developing the field of IRE, beginning with the current revision of the Association of Internet Researchers' (AoIR) ethical guidelines. These guidelines, adopted by AoIR in 2002, have found extensive use around the world as a helpful guide to analyzing and resolving ethical issues in Internet research. The current revision seeks to update the guidelines in light of the dramatic expansion of Internet research following on the emergence of so-called Web 2.0 technologies and the ongoing global diffusion of the Internet. The second project is the Internet Research Ethics Digital Library, Research Center, and Commons (<http://www.internetresearchethics.org/>). This ongoing project is the result of a grant awarded by the National Science Foundation to the Center for Information Policy Research at the University of Wisconsin-Milwaukee's School of Information Studies. A primary goal of this project is to develop and provide sound resources, a solidified research base, and expert advice as more researchers and more IRBs/ethics boards struggle with the complexities of Internet research ethics. Both projects thus share an emphasis on praxis – i.e., analyzing and responding to real-world dilemmas faced by a growing research community around the globe.

Following these introductions and overviews, the panel will invite critical discussion of the representative issue, approaches, and resources. As well, the panel will welcome comments and suggestions from participants for additional resources and insights that will contribute to both projects – and to suggest ways where these projects in turn contribute to contemporary work in information and computing ethics. A last goal of the panel is to develop a better articulation – a conceptual map – of the multiple relationships between IRE as a field of information and computing ethics and other characteristic foci and thematics of computing and philosophy.

# **Track I: Philosophy of Computer Science**

## RULES AND PROGRAMMING LANGUAGES

RAINHARD Z. BENGEZ

*Philosophy of Science, Technology, and Engineering Department*

*Carl von Linde Academy*

*TUM School of Education*

*TU München, Arcisstr. 21, 80333 München, Germany*

*[bengez@tum.de](mailto:bengez@tum.de)*

### Abstract

In computer science and related fields we are talking much about rules. The word *rule* appears very often directly or unspoken in papers concerning computer science or Philosophy of Computer Science. We are talking about logic(s), interpreters, procedures and compilers, systems of rules, programming languages, automata and rules of software design, good practices, and much, much more. But, unfortunately, the meanings of the word *rule* to which one refers from case to case seem to be unclear. In my contribution I would like to try to show some of these ambiguities and discuss ways to avoid them. According to the nature of this subject, my contribution is both analytical and normative as well, because I will analyze some applications of the word and work out a traceable direction for use of it. Admittedly, the word *rule* has so many directions for use in computer science and philosophy of computer science that I cannot talk about most of them. I will restrict myself to rules inducing action and especially to such rules in programming languages (DSL, specification, etc.). This would mean rules are guiding actions in languages, or, stated more general, in sequentially structured patterns. I will start by talking about the dependence of rules and actions.

## A BEHAVIORAL CHARACTERIZATION OF COMPUTATIONAL SYSTEMS

JAVIER BLANCO

*Universidad Nacional de Cordoba, Argentina*

RENATO CHERINI

*Universidad Nacional de Cordoba, Argentina*

MARTIN DILLER

*Universidad Nacional de Cordoba, Argentina*

AND

PÍO GARCÍA

*Universidad Nacional de Cordoba, Argentina*

**Abstract.** We introduce the concept of interpreter as a producer of behavior in response to some input that codifies it. We argue that the notion of interpreter captures the minimal characteristics shared by different kinds of computational devices, and can thus serve as a criteria to identify how interesting a computational system is. This characterization contrasts with many of the current functional descriptions offered in the literature on this topic, in that these are somewhat dependent on the technology that is currently available. Since the concept of interpreter can be used to compare different systems, it defines a computational hierarchy, establishing the relative degree of computationalism of different systems. This enables us to restate some ontological questions, such as what is a program?, when is a system computational?, in more precise terms which admit clearer answers.

Any system can be characterized in terms of its possible behaviors. In particular, a useful description of a computational system is given by the relationship between the input and the behavior produced as a response to that input, characteristic of the system.

The feature that distinguishes computational systems from other types of systems is that they may produce a very large and interesting set of behaviors, depending on syntactic inputs and “without changing a single wire” (Dijkstra, 1988). Thus, the characteristic input-behavior relation implicitly defines an encoding of behaviors as syntactic objects.

We have suggested in (Blanco et al, 2011) that some key aspects of computational systems can be captured by the ubiquitous concept of interpreter as used both in

theoretical and applied computer science (Jones 1997, Abelson&Sussman 1996, Jifeng & Hoare, 1988), defined in a very general manner. In this article, we present an interpreter as the necessary link between a set of behaviors and their respective encodings, without relying on any mechanistic account of systems. As we argue elsewhere, the concept of interpreter can be regarded not only as a notion that captures the minimal common characteristics of different types of computational devices and serves to clarify various concepts which pervade computer science, but also as a framework for understanding computing.

By *behavior* of a system we understand only a description of the occurrences of certain events considered relevant of the system. Different ways of observing a system may determine different sets of behaviors. Thus, the behaviors will depend on a decision regarding the events that are considered of interest for that system (for particular purposes). A precise definition of behavior will be left unspecified here, since this will only make sense when a particular framework is stated.

Intuitively, an interpreter produces a behavior according to some input that codifies it. Usually, the encoded behavior may depend on input data, but for simplicity we will assume in this presentation that the data and behavior are already encoded together. The notion of interpreter is (almost) by definition the necessary link between the so-called “program-scripts” and “program-processes” (Eden 2007, Blanco & Garcia 2008).

Given a characterization of a fixed set  $B$  of possible behaviors, and a set of syntactic elements  $P$ , an *interpreter* is a function  $i : P \rightarrow B$  assigning some behavior  $b$  in  $B$  to every  $p$  in  $P$ . When this relation is given we say that  $p$  is the *encoding* of  $b$ . Generally, we speak of the syntactic domain  $P$  as the *programming language*, and of  $p$  as a *program*.

A (physical) system  $I$  realizes an interpreter  $i$  if it is capable of receiving an input  $p$ , and systematically produce the observable behavior  $b$  such that  $i(p) = b$ . In this case we say that  $I$  *effectively computes*  $b$  via the program  $p$ . We say that a (physical) system *realizes* an interpreter when every time we provide it with an instance of an encoding, it produces the corresponding observable behaviour. We do not consider internal states, since these may be realized in very different ways.

One way of precisising the notion of realization is along the lines of the notion of “practical realization of a function” defined in (Scheutz 1999), where the relation is an isomorphism between the formal definition of  $i$  and a physical theory  $T$  that describes the system  $I$  (for example, the theory of electrical circuits) that includes a description of the inputs and outputs of the system as well as a function  $F$  that maps inputs to outputs using the laws and language of  $T$  in a way that guarantees the preservation of the isomorphism. In (Scheutz 1999) different degrees of “practicality” of the realization relation are also considered that take in account the limits in precision with which the inputs can be measured and generated, reliability and range of functioning of physical systems, noise generated by the environment, etc.

The concept of interpreter serves as a criteria to distinguish between systems that could be computational (w.r.t some inputs and behaviours) from those that could not. Since we want to capture what makes any system programmable, we do not assume any particular implementation technology in the concept of interpreter. Different computational models, like Von Neumann machines, parallel machines, DNA-computers, quantum-computers, can be considered interpreters because they can systematically produce behaviours from their encodings in a predefined language. What

will be specific to each model is the underlying theory used to justify that they are interpreters, not the criteria used to determine that they are indeed programmable systems.

The notion of interpreter can be seen as functional, i.e, an interpreter is such when it is capable of producing behaviors from programs. Following this idea, a program is a syntactic structure capable of being interpreted. A program is such only relative to a given interpreter and an interpreter is such only for a particular programming language. The concepts of program, programming language and interpreter are thus relational and inter-definable.

The main feature of an interpreter is that it is *programmable*: there is an available syntax with which a variety of behaviors can be encoded. The degree of programability of an interpreter is given by the variety of behaviors that the underlying programming language is able to encode. The *degree of programability* is the distinctive feature of an *interesting computational system*. If we consider a system computational when it is programmable, then being computational will also be a property which can be established only relative to a set of behaviors and a corresponding encoding (usually an actual programming language). In other words, the property of being computational will not make sense independently from a set of behaviors and the encoding. This will allow us to tackle some philosophical problem such as the problem of pan-computationalism (do all physical systems compute?) (Putnam 1987, Searle 1990, Chalmers 1996, Chrisley 1994, Copeland 1996, Piccinini 2008) from a different perspective. The question “Is this a computational system?” is replaced by the question “Is this a computational system with respect to this set of inputs and behaviors?”, or equivalently, “How interesting, from a computational point of view, is this system?”. From this perspective, in particular, several constructions of “trivial implementations of programs” which intend to show how the thesis of pan-computationalism can be established do not qualify as interesting computational system.

Since the rise of computability theory in the thirties, it was clear that a computation is related to certain formal object that prescribes it, e.g. the description of a Turing Machine, general recursive functions, a lambda-term, etc. A computation, then, is produced following this prescription. Putnam’s (and Searle’s) theorem (Putnam 1987, Searle 1990), on the other hand, tries to present a notion of computation in itself, reifying computation as something that exists independently of the prescription or program (any sequence of states would do).

The property of being an interpreter for a given set of behaviours can be satisfied by certain systems. An interpreter is a general notion that can be used to characterize physical mechanisms (computers, calculators), a human acting mechanically (Turing’s computer, a human carrying out the reductions of a lambda term), mathematical formalisms (universal Turing machines, etc.), or computers with computing power beyond Turing computability (Oracle computers (Copeland 2002)). Whereas a (physical) counterpart is needed for the realization of an interpreter, the property of being an interpreter, and concomitantly, the property of being a programmable system, can be determined by its abstract description.

## References

- Abelson, H. & Sussman, G.(1996) *Structure and Interpretation of Computer Programs*. MIT Press, Cambridge, MA, USA, 2<sup>nd</sup> edition.
- Blanco, J., Cherini, R., Diller, M. & Garcia, P. (2011) *Interpreters: towards a philosophical account of computer science*. Technical Report.
- Blanco, J. & Garcia, P. (2008) A categorial mistake in the formal verification debate. In *European Conference on Computing and Philosophy (ECAP)*, June 2008.
- Chalmers, D. (1996) Does a rock implement every finite-state automaton *Synthese* 108 (3):309-33.
- Chrisley, R..(1994) Why everything doesn't realize every computation. *Minds and Machines*, 4(4):403-20
- Copeland, J.(1996). What is computation? *Synthese*, 108(3):335-59,
- Copeland, J.(2002) *Narrow versus wide mechanism*. In *Computationalism: New Directions*. MIT Press.
- Dijkstra, E..(1988) *On the cruelty of really teaching computing science*. circulated privately.
- Eden, A..(2007) Three paradigms of computer science. *Minds Mach.*, 17(2):135-167.
- Jifeng He. & Hoare, C. (1988) Unifying theories of programming. In Ewa Orłowska and Andrzej Szalas, editors, *RelMiCS*, pages 97-99.
- Jones, N. (1997) *Computability and complexity: from a programming perspective*. MIT Press, Cambridge, MA, USA.
- Piccinini, G (2008) Computers. *Pacific Philosophical Quarterly*,89(1):32-73.
- Putnam, .H.(1987) *Representation and Reality*. MIT Press.
- Scheutz, M (1999). When physical systems realize functions. *Minds and Machines*, 9(2):161-196.
- Searle, J (1990). Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association* 64 (November):21-37.

## WHAT IS THE DIFFERENCE BETWEEN YOUR FRIEND AND A CHURCH-TURING LOVER?

*A New Defense of H-Consciousness.*

PIOTR BOŁTUĆ  
*University of Illinois Springfield*  
*UHB 3030, One University Plaza*  
*Springfield IL 62703*  
*(and Warsaw School of Economics)*

**Abstract.** Whatever functionality may be attained by a physical system, (such as a human), it could, be replicated by a robot. We can define a Church-Turing lover as a robot with all functionalities of a (realistic, or ideal) sex partner. What it lacks is only the first person perspective. If we care what a partner truly feels, not just how he/she behaves, we should care. Yet, if we could build-in relevant first-person consciousness, the difference would disappear, or it would be relegated to a broader social-historical context..

### 1. The gist of the Argument

An important direct implication of the Church-Turing seems to be that whatever functionality may be attained (by a physical system, such as a human), it can, in principle, be replicated by a robot. In the area of sex, whatever ‘functionalities’ a human lover may perform, the same would in principle be replicable in advance sex-toys. The term ‘functionality’ can be understood as broadly as we can. Should desired specifications of a lover include, in addition to advanced mechanical functionalities, also certain advanced tactile features, temperature adjustments, fluid emissions (including chemical replication of the body fluids, such as sweat, squirt or sperm), ionization levels and other bioelectrical fields, sounds or even sophisticated conversations and other language utterances ( ‘the Turing test’ is one of the implications of Church-Turing) such conditions can be produced, though sometimes the cost may in practice be prohibitive. To understand this point is important for the large sex-toy industry, for other industries piggybacking on its research and development, but also for the philosophers. The question for philosophers is what, if anything, would make such robotic lover different from a human one. Advanced robotic lovers can be viewed as external experience machines, where one’s senses are stimulated by an artificial cause but not through direct brain stimulation but rather through stimulation of external sensory organs. It is however similar to the experience machine since the robot breaks the ‘typical’ or so-called

‘proper’ causal chain between the experiences and a human lover and initiates a so-called deviant causal chain (terms ‘proper’ or ‘deviant’ are used here in the sense used in theory of causality, not as moral evaluatives). I come to the conclusion that, while there is no functional difference, the human lover is supposed to have a first-person (h-consciousness) related to Chalmers’ hard-problem.

Without such assumption we have no way to philosophically articulate the difference between the moral subjects for whom ‘there is something that it is like to experience’ a certain thing (here, sex) for the inside, and those for whom there isn’t such a thing. Perfect electronic lovers work better than zombies in demonstrating this point since we avoid the controversies whether it is conceivable that identical physical systems, such as human brains, could produce first-person consciousness in humans but not in zombies. The zombies seem to violate the tenet of materialism that there is no difference without physical difference while electronic toys do not make such violations..

### 1.1. MAIN STEPS OF THE ARGUMENT

Let us present a ‘sentence outline’ of the main argument.

#### *1.1.1. Defining a Church-Turing lover*

It is the perfect functional imitation of a human lover in terms of all parameters desired, which may include some or all of the following: a. tactile features, b. reactivity to voice commands, c. speech quality, d. speech content (including, the ability to meet the Turing test), e. advanced domestic skills (cooking, cleaning), f. other skills of an artificial companion as defined by Floridi.

#### *1.1.2. Defining your boyfriend/girlfriend*

Defining your boyfriend/girlfriend as a human being, equal or inferior to the Church-Turing lover in terms of the functionalities described broadly in points a-f and all other typical functionalities.

#### *1.1.3. Establishing rough functional equality between the Church-Turing lover and the Boyfriend/Girlfriend.*

This includes the responses to various objections such as the social objection, the psychological objections and the religious objection. The only objection left unanswered is the reproductive objection, which leaves us with ‘rough functional equality’: Church-Turing lover is functionally equal to your Boyfriend/Girlfriend provided you do not intend to procreate with him/her. (Actually, Church-Turing implies procreative functionality in robots as well).

#### *1.1.4. Atypical functionalities, defined as those of the first-person perspective.*

I show futility of the Church-Turing functional reenactments of presumed first-person states. *Why do I want my boyfriend/girlfriend to have an orgasm not just to be very good at faking one?* (If I am not an egoist I want her to feel good not just to behave as if she felt so.) Also, I give a brief, responses to the privileged access problem).

*1.1.2. The engineering thesis in machine consciousness*

The engineering thesis in machine consciousness, saves your girlfriend/boyfriend's uniqueness, but not forever. There is a first-person, inductively established, difference between the Church-Turing lover and a boyfriend/girlfriend. The difference may partially disappear should we be able to engineer robots with first-person h-consciousness.functionalities.

*Acknowledgements*

I developed an early draft of this argument at G. Harman's graduate seminar in epistemology in the Spring of 1991. I want to thank Prof. Harman, Alex Byrne, Mary McGowan and other participants for discussion. I want to thank John Barker and Keith Miller for recent related discussions.

## HAECCEITY AND INFORMATION

THEPTAWEE CHOKVASIN  
*Suranaree University of Technology*  
*Nakhon Ratchasima, Thailand*

**Abstract.** The interest in ‘information entities’ is increasing in the philosophy of information. In this article, I offer a philosophical analysis which is concerned only with their haecceities (thisnesses) in the conception of Heideggerian ‘functionality’. I argue that the haecceity of an information entity is necessary for making a legal judgment on cybercrimes- especially on sharing illegal information. Moreover, when considering about the persistence of deleted information files, it is found that haecceities of those information files have some aspect of being an indexical of functionality which is far beyond what Duns Scotus knew about them.

### 1. Introduction

I live in Thailand, and my friend is now in Japan. We are chatting on the MSN. If now I’m reading some information in a school website, and my friend is reading the same thing on his computer screen in Japan, are we exactly reading *the same* thing?

Someone may consider about this situation and say that the same thing can appear in many different places at the same time, therefore we are exactly reading the same thing. However, some other may say that one thing cannot be in many different places at the same time, so my friend and I are looking at two different website pages which are merely similar to each other.

And so, a question arises, “*When are two chunks of information, or two information entities, the same?*.” In this fashion of the argument above, it can be seen that something that is very similar to the problem of universals is brought back from classic metaphysics. Cyber-information on webpage behaves like it is a universal which is instantiated in many individual computers. However, if a philosopher of information wants to retain the position of considering information as information entities, she may have to take another route of explaining the similarity of the two web-pages. She might explain that they are two different information entities that instantiates the same universal ‘informativeness’.

If the latter is right, then we have to admit that any information is an information entity of its own. There are no two distinct information entities exactly resemble each other. Unfortunately, this position of metaphysical information entities may have undesirable result. In the present time, there is a law of computer crime that forbids sending or forwarding any illegal information, pictures, piracy items, etc. to a third

person. Both of the sender and the receiver will be considered guilty of doing that. But how can the law still be legitimate if the receiver uses the argument above to show that because of their status of being different information entities, he therefore did not receive *the same* thing from the sender?

The latter one leads us to other topics in metaphysics which are about identity and individuation, and in this article it interests me more than to find out the account of sameness of information entities in the light of the metaphysics of universals. So, I will stick to the topic of identity and individuation. In this article, I will develop an analysis to answer the question above. The analysis will be in the light of Heideggerian 'functionality' as mentioned by Ratcliffe (2002) that, apart from their properties, for two things to be identical to each other they must be considered from their 'teleological webs' including their values and ends. However, it must be developed further when answering another question of what the appropriate notion of identity for information entities is. I will argue that the problem of individuation is deeper than the problem of identity. The two information entities that are not different in their properties will be individuated by their info-haecceities which are the bases for their identity.

## 2. Haecceity and Functionality

It is said that John Duns Scotus may be the first philosopher who deals with the problem of individuation with "the difference". Duns Scotus gave arguments for positing an "individuating difference" or a haecceity which is to give an account to individuals. In his *Ordinatio*, Duns Scotus said that "I reply therefore to the question that material substance is determined to this singularity by some positive entity and to other diverse singularities by other diverse positive singularities." (Wolter, 1994 : 286).

The positive individuating difference, or haecceity, is different from the common nature, or quiddity, that is to explain *what* an individual essentially is. So, we may never reach a full understanding of the haecceity.

Now we can say that the receiver of the illegal information may be considered guilty from another perspective. Although it can be said that it is controversial of him being guilty of receiving the very same thing from the sender, he is still guilty from producing another new illegal entities in the computer system. It has to depend instead on "the difference" to be legitimate for charging to two persons (not just one) of being guilty of two different acts differentiated by two different entities which just happen to have the similar characteristics in their common natures.

Cannot haecceity be grasped at all? In *Haecceity* (1993), Gary S. Rosenkrantz had some arguments to show that the haecceity of the objects incapable of consciousness are to us cognitively inaccessible. Only the haecceity of one's being oneself can be grasped and expressed linguistically by only that one person. If we follow Rosenkrantz's argument, we have to admit that the haecceity of other entities around us is inaccessible. Is this the same case for haecceity of information entity, or info-haecceity?.

## References

- Ratcliffe, Matthew. (2002). Heidegger, Analytic Metaphysics, and the Being of Beings. *Inquiry* 45(1), 35-57.
- Rosenkrantz, G. A. (1993). *Haecceity: An Ontological Essay*. Dordrecht: Kluwer Academic Publishers.
- Wolter, A. B. (1994). John Duns Scotus. In Jorge J. E. Gracia (ed.), *Individuation in Scholasticism: The Later Middle Ages and the Counter-Reformation 1150-1650* (pp. 271-298). Albany, NY: State University of New York Press.

## THE LIMITS OF COMPUTER SIMULATIONS AS EPISTEMIC TOOLS

JUAN M. DURAN  
*Universität Stuttgart - SimTech*  
*Germany*

Over the past few decades the use of computers for scientific purposes has been extended to virtually every branch of science. Such widespread acceptance is clear: their provide powerful means for solving complex models, as well as speed and memory for analyzing and storing data, visualizing results, etc.

A less broad, yet still important, use of computers in laboratory practice is by means of implementing computer simulations. Lately, scientists have turned their interest to the design, validation, and execution of computer simulations instead of setting up, controlling and calibrating a whole material experiment. Whether for budgetary reasons, time-consuming delays, or complexity, today scientific practice is carried out in a way that strongly relies (if not fully depends) on computers. Here we face a philosophical problem that now has become widely discussed.

Current philosophical literature deals with the question whether the epistemological value of a traditional experiment has greater (or less) confidence than a computer simulation. The most used trick for answering this question is by addressing the so-called “materiality problem”.

Its standard conceptualization is characterized by Parker in the following way: “in genuine experiments, the same ‘material’ causes are at work in the experimental and target systems, while in simulations there is merely formal correspondence between the simulating and target systems (...) inferences about target systems are more justified when experimental and target systems are made of the ‘same stuff’ than when they are made of different materials (as is the case in computer experiments)” (Parker, 282). In general terms, the materiality problem can be addressed either by emphasizing the lack of materiality in computer simulations as epistemically defective (for example, as in Guala, Morgan and Giere), or by claiming that the presence of materiality in experiments is rare and, ultimately, unimportant for epistemic purposes (Morrison, Parker and Winsberg).

Either solution leads to what I call the “dilemma of computer simulations” for it presupposes that once the ontology of computer simulations is sorted out, its epistemic power can be fully determined. Indeed it is required, as premise, to provide an ontology that resolves the epistemic value of computer simulations. However, the informative exercise of simply checking off ontological features of computer simulations begs the question whether it is legitimate to draw any epistemic conclusion at all. Paraphrasing Hacking, they disagree because they agree on basics.

A different approach consists of defending the epistemic reliability of computer simulations as philosophically detached from its ontological conceptualization. This does not suggest, though, that they are two unrelated issues, but instead that each can be analyzed in its own right. In fact, there exist a close relation between them insofar the ontology becomes, to certain extent, a limiting case for the epistemology of computer simulations.

Therefore, instead of asserting that “on grounds of inference, experiment remains the preferable mode of enquiry because ontological equivalence provides epistemological power” (Morgan, 326), I hold a twofold claim: firstly, that materiality only restricts computer simulation from “accessing” certain aspects of the world which require a causal story; in other words, materiality draws the boundaries from where experiments become a specific and irreplaceable method for knowing something about the world. Secondly, that computer simulations provide ways of inference that do not depend on its materiality but on its capacity for representing empirical as well as non-empirical systems. □□ Keeping an eye on these two claims, I propose to proceed in to correlated steps: firstly, by analyzing and characterizing the nature of computer simulations and material experiments; naturally, this step is highly dependent on assumptions on computational models, computer programs and experiment, all of which will be briefly addressed. Secondly, by discussing the philosophical relevance of the limits imposed to computer simulations by materiality as well as drawing some preliminary conclusions on their epistemic power.

Case examples will be briefly discussed as well. In one sense, there are many aspects of scientific practice that cannot be substituted by computer simulations, but require interaction with the material world: measurement, for instance, is one case. In certain measurement instances (i.e. the so-called “derived measurement”), the causal interaction of an instrument with the world cannot be replaced by the calculus performed by a computer simulation. Another interesting case-study is the reproducibility of experiments (Cf. Franklin and Howson 1984): as it is well known, the variation of instruments and experimental set-up tends to increase its epistemic reliability; it is not clear, however, that a similar methodology may work for computer simulations. In addition, the detection of new real-world entities seems a complete chimera for computer simulations, although it is a key role of material experiments. On the other hand computer simulations have the capacity of dealing with incredible complex equations that represent real-world systems and from which it is possible to “crunch” large amounts of data. Most of our knowledge about the world also comes from manipulating and interpreting such data. Computer simulations can also be used for investigating “rational worlds”, such as counterfactuals, thought experiments and mathematical worlds.

I then urge for a philosophical discussion of the epistemological value of computer simulations based on its capacities and limits, instead of the dependence on an ontological conceptualization.

## References

- Franklin A., and Howson, C. (1984), Why do scientists prefer to vary their experiments?, *Studies in History and Philosophy of Science Part A*, 15(1), 51 – 62.

- Giere, R. (2009), Is computer simulation changing the face of experimentation? *Philosophical Studies*, 143, 59–62.
- Guala, F. (2002), Models, simulations, and experiments. In: L. Magnani and N. J. Nersessian (Eds), *Model-Based Reasoning: Science, Technology, Values* (pp. 59-74). Kluwer.
- Morgan, M. (2005), Experiments versus models: New phenomena, inference and surprise. *Journal of Economic Methodology*, 12(2), 317–329.
- Morrison, M (2009), Models, measurement and computer simulation: the changing face of experimentation, *Philosophical Studies*, 143, 33–47.
- Parker, W. (2009), Does matter really matter? computer simulations, experiments, and materiality, *Synthese*, 169(3), 483–496.
- Winsberg, E. (2009), A tale of two methods, *Synthese*, 169(3), 575–592.

## WHY TO BUILD A PHYSICAL MODEL OF HYPERCOMPUTATION?

FLORENT FRANCHETTE

*IHPST, University of Paris 1 Panthéon-Sorbonne  
13 rue Dufour, 75006 Paris*

**Abstract.** A model of hypercomputation can compute at least one function not computable by Turing Machine and its power comes from the absence of particular restrictions on the computation. Nowadays, some researchers claim that it is possible to build a physical model of hypercomputation called “accelerating Turing Machine”. But for what purposes these researchers would try to build a physical model of hypercomputation when they already have mathematical models more powerful than the Turing Machine? In my opinion, the computational gain provided to the accelerating Turing Machine is not free. This model also lost the possibility for a human to access to the computation result. To define this feature, I will propose a new constraint called the “access constraint” stating that a human can access to the computation result regardless of computation resources. I will show that the Turing Machine meets this constraint unlike the accelerating Turing Machine and I will defend that build a physical model of the latter is the solution to meet the access constraint.

The aim of the computability theory is to define mathematical functions computable by algorithms. The definition of an algorithm is however an informal one and the computability theory needs a mathematical definition of this notion. In order to formalize a predicate which means “can be computed by an algorithm”, Alan Turing (1936) proposed the formal predicate of “computed by Turing Machine” or “Turing-computable”. According to Turing, the Turing Machine (TM) is a mathematical model of computation with a power equivalent to an algorithm. This claim is summarized in the Church-Turing thesis: functions computable by algorithms are computable by TM. This thesis argues that the TM defines the computation by algorithm since if a function is not Turing-computable, there is no algorithm which can compute it. For example, Turing proved that some mathematical functions such as the Diophantine function<sup>1</sup> are not Turing-computable. Turing (1939) however, showed in his thesis that the computing power of the TM, that is to say the number of functions it could compute, depended on the type of constraints applied to the model.

Models which are able to compute more functions than the TM are called “models of hypercomputation” or “hyperMachine”, and their computational power comes from

---

<sup>1</sup> Given a Diophantine equation  $x$ , the Diophantine function is the function such as  $f(x)=1$  if  $x$  has at least a solution and  $f(x)=0$  otherwise.

the absence of particular restrictions on the computation. Recently, Jack Copeland (2002) has proposed a model of hypercomputation named “Accelerating Turing Machine” (ATM) which is based on the absence of the constraint that the computation must include a finite number of steps. Copeland demonstrates in his article that an ATM is able to execute an infinite number of computational steps in a finite time and compute non Turing-computable functions such as the Diophantine function. More importantly, some researchers defend the idea that it is possible to physically build an ATM. However, the physical construction of a computational model, whether equivalent to the TM or not, goes beyond the original framework of the computability theory. Indeed, the Church-Turing thesis states nothing about the computing power of a TM physically built, it states only an equivalence between the intuitive concept of algorithm and the mathematical concept of Turing Machine. It is therefore pertinent to ask for what purposes these researchers would try to build physical hyperMachines when they already have mathematical models more powerful than the TM. In other words, why leave the mathematical framework of hypercomputation to turn to the physical sciences?

In order to answer these questions, I will try to explain one reason why advocates of hypercomputation want to physically build a computational model with a greater power than the TM. In my opinion, although the absence of a constraint such as the finite number of steps allows the ATM to compute more functions than the TM, the computational gain is not free. The model of hypercomputation also lost a key feature: the possibility for a human to access to the computation result. To define this feature, I propose a distinction between “to access to the result” and “to compute the result”.

We have access to the computation result when the result is available to us in principle. This result doesn't need to have a meaning, it can only be a string of symbols.

We compute a result when we can follow in principle each computational step from input to output.

From these definitions, we can set out two constraints: one asserting that we can compute results computed by a model and the other asserting that we can have access to these. Let a function  $f$  which is computable by a model.

- This model meets the access constraint (AC) if for all input  $x$ , we can have access to  $f(x)$ .
- This model meets the computing constraint (CC) if for all input  $x$ , we can compute  $f(x)$ .

It is straightforward to show that these two constraints are set out in the definition of a TM. However, I think that the ATM doesn't meet the CC and the AC. My main point is to explain that it is actually unlikely that a human can compute an infinite number of steps in a finite time. This argument consists to say that the brain, where computations are made, is a finite entity both in space and time. This argument seems pertinent in order to show that we are not able to follow step by step an infinite computation. But it is not sufficient to prove that we can't have access to the result from an infinite computation because it could be possible that we have access to Diophantine function results without to follow each computational step. For example, Hava Siegelmann (1995) has proposed a mathematical model of the brain in the form of artificial neural nets which according to her could compute “beyond the Turing limit” Although it appears that Siegelmann's model may exceed the power of the TM, it has been strongly criticized by Martin Davis (2006) in his article entitled *The myth of hypercomputation*.

From the two arguments outlined above, I shall make the assumption that a human is not able to compute and to have access to the result of a non Turing-computable function

computed by an ATM. Therefore, this model does not meet the CC and the AC. Nevertheless, could an ATM meet these constraints? In my opinion, it is necessary to distinguish two ways for a model to meet the AC.

- A model meets the AC in an internal sense if a human is able to have access to the computation result without a physical realization of the model.
- A model meets the AC in an external sense if a human is able to have access to the computation result with a physical realization of the model.

For example, a TM meets the AC in an internal sense because we can access to results from its mathematical definition. On the hypercomputation side however, we could have access to the computation result in an external sense with a physical realization of an ATM. This result, characterized by the link between the computing power of a model of hypercomputation and its physical realization has important consequences for the notion of computation. It shows that some features belonging to hypercomputation models do not only depend on mathematics. Specifically, the possibility to access to the result of a non Turing-computable function computed by an ATM is based on physical constraints.

### Acknowledgements

I would like to thank the editors and referees for very helpful comments during the preparation of this paper.

### References

- Copeland, J. (2002). Accelerating Turing Machine, *Minds and Machines*, 11, 281-301.
- Davis, M. (2006). The Myth of Hypercomputation. In C. Teuscher (ed), *Alan Turing: the Life and Legacy of a Great Thinker*, Springer.
- Turing, A. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem, *Proceedings of the Mathematical Society*, 42, 230-265.
- Turing, A. (1939). Systems of Logic Based on the Ordinals, *Proceedings of the Mathematical Society*, 45,161-228.
- Siegelmann, A. (1995). Computation Beyond the Turing Limit, *Science*, 268, 545-548.

## THE MATERIALISTIC FALLACY

### *Some Ontologic Problems of Regulating Virtual Reality*

FABIAN GEIER  
*Universität Bamberg*  
*An der Universität 2*  
*96047 Bamberg*  
*Germany*

**Abstract.** This paper will discuss a connection between the ontology of virtual objects and several problems of information ethics. I argue that there is a strong tendency, sometimes even among professionals in ICT, to treat virtual objects like material objects. There are many political regulations and economic practices which make sense for material objects, but do not make sense for virtual ones. Such an ignoring of the nature of data processing, be it deliberate or not, I call a materialistic fallacy and consider it to be hampering social progress and benefit.

### **1. The Fallacy**

I call a materialistic fallacy if virtual objects are unnecessarily treated like material objects. The immediate effects of this fallacy are two: The practice in question either proves to be ineffective, because it is easily circumvented; or, where it can be enforced, it stalls progress and severely limits the benefit that ICT could provide.

### **2. The Ontology of Virtual Objects**

By “virtual objects” I refer to any chunk of digitally stored data that is conceived as a distinct entity by human understanding. This will in most cases be identical with files. However, the human mind does not have to go along the lines of file descriptors, and especially outside professional IT it often does not. A mouse pointer, window or web-page might be made up of several distinct files, and neither is a part of a file a file, nor is the entire content of a hard-drive. However, all these are virtual objects, as soon as we refer to them. And the decisive thing about virtual objects is that they can easily be made a file and be subject to all possibilities of data processing. By this definition of virtual objects I hope to circumvent most of the specific problems in the ontology of computing.

In material reality, form and matter cannot be separated from each other. One of the effects of this is, that we are used to relatively stable individual objects, that persist in time. Persistence is the precondition for movement: When a material object is moved into a new place, it is not at the same time in its former place anymore.

In the realm of information, however, the case is entirely different. In Aristotelian terms data processing deals with pure 'forms'. Forms don't move. They are a-temporal and intangible (This largely corresponds to what Eden & Turner (2007) say about programs). Their distinctive characteristic is *instantiation*. Any number of instances of a form can exist, but none of them is prior to any other. If we send a network packet to two different computers, we cannot say which of the arriving packets is the original and which a mere copy. Such questions make sense in the material world, but they do not make sense in the virtual.

Technically, any chunk of data is at any point located in particular bits and bytes, and so still is an instantiation and not a pure form. However, since computers are all about reinstantiating the form of this instantiation, this fact is negligible. Computers are all about *making* it negligible. This results in what Moor (1997) calls information being "greased".

Of course there seems to be movement in virtual objects, i.e. in a cursor on a screen. Otherwise computers would not be very useful. But we should keep in mind that such movement is always a *simulation*, created by a sequence of copying and erasing. But only because we sometimes cannot help using such simulations, there is no need to do it to the utmost degree. I suggest the opposite: We should do it only where it is necessary, and otherwise maximize the benefit from freeing information from the bonds of materiality.

### 3. Examples

#### 3.1 DATA EXPIRY

A typical materialistic fallacy is the suggestion, put forward by Viktor Mayer-Schönberger (2008), and recently picked up by the German ministry for consumer protection, to have an inbuilt expiry date for data on the internet. The idea sounds nice: This would end the problem, that what is put online once, resides there forever.

However, it will never work. More precisely: It could only work under the most extreme conditions of worldwide data-control – an amount of control no current institution is anywhere close to exert. Of course we can write a program that erases a file after 90 days, but it would have to be implemented either as a mandatory core module of all existing operating systems, or as an obligatory hardware solution similar to Trusted Computing. However, it does not lie in the nature of data to expire. An expiry module would only be a separate addition to the core functionality of computers, and thus both unwanted and easy to remove.

#### 3.2 DIGITAL RIGHTS MANAGEMENT

DRM, or more specifically, copy protection is almost archetypical for the materialistic fallacy. When we are trying to charge customers on a per-copy basis, we are following

the paradigm of material objects. Copy protection attempts to establish a uniqueness and sameness for the copy, that does not lie in its nature. The protection must prevent a function that data processing generally offers: the re-instantiation of data.

There are various consequences of this: First, the moral restraints to copy software, protected or not, are lower than in material theft, because copying does not result in anyone else losing data. Second, just because it is not its nature, the seeming uniqueness of a copy is difficult to maintain, as it can only be provided by an additional module. I do not endorse pirating software. But I endorse acknowledging the basic structures of ICT because of which it is easier to pirate it than to protect it. And I endorse thinking about alternative ways of dealing with this.

### 3.3 E-VOTING

The ontologic structure of ICT also matters in the discussion about eVoting. I am not referring to security issues here, but to the situation once security is breached. Then the full power of data processing lies at the hand of the intruder: Whether you forge 10 votes or 10 000 000 – it is just one line of code. The difference between local and global modification is not the same in virtual as in material reality. Virtual Objects do not count one by one, but can be treated formally, on various levels of abstraction. Large scale modifications in a database are in principle no more difficult than singular modifications. I don't say that this alone must decide the issue. All I say is that the nature of data processing has to be taken into account.

## References

- Aristotle (1998). *Kategorien. Hermeneutik*. Hamburg: Meiner Verlag.
- Aristotle (1989). *Metaphysik*, 2 Volumes. Hamburg: Meiner Verlag.
- Mayer-Schönberger, V. (2008). Nützliches Vergessen. In: M. Reiter, M. Wittmann-Tiwald, (Eds), *Goodbye Privacy - Grundrechte in der digitalen Welt*. Wien: Linde Verlag.
- Moor, J. (1997). Towards a Theory of Privacy in the Information Age. *Computers and Society* 27 (3), 27-32.
- Eden, A. H. & Turner R. (2007). Problems in the Ontology of Computer Programs. *Applied Ontology* 2 (1), 13-36.

## THE EFFECT OF COMPUTERS ON UNDERSTANDING TRUTH

STEVEN MEYER

*Tachyon Design Automation Corp.*

*Minneapolis, MN*

**Abstract.** The effect of computers and computation on the philosophical study of the epistemology of truth is discussed. The development of algorithmic truth as satisfiability is considered using modern quasi empirical methods that follow the mathematician Paul Finsler's discovery that a formal conception of truth does not suffice. The P=?NP problem is considered and shown to be a philosophical problem using Finsler's method. Non truth value assignment conceptions of truth such as deflation and computer science as a method for studying physics are criticized.

### 1. Introduction

The mid 1960s marked the beginning of the influence of computers on the epistemology of various conceptions of truth. On the one hand fast computers were becoming available and on the other quasi-empirical characterizations of mathematics in the form of Lakatosian research programmes were becoming popular (Lakatos, 1967). A. J. Ayer attributes the quasi-empirical characterization of logical truth to J. S. Mill from the middle of the 19th Century (Ayer, 1936, p. 291). In 1964, Paul Finsler published what he claimed was an air tight defense of his rejected 1926 idea that 'A Formal "conception of truth" cannot suffice' (Finsler, 1996, p. 163).

Computers were becoming fast enough so that computer programs for proving mathematical theorems and for verifying truth were conceivable. These developments led naturally to questions concerning what can be computed, and if there are any limitation of computability. Before the mid 1960s, at least in the area of mathematics, epistemology had become truth as existence of mathematical objects generated from abstract set theory. The various incompleteness, inconsistency and set theory paradox results were avoided by falling back on truth as axiomatic logic.

Computers allow a new and seemingly empirical epistemology of truth. Namely, something is true if it can be computed in a reasonable amount of time. This immediately led to problems. One early example was alphabetization (sorting) using a giant table. One can sort a list in linear time by converting each key into a number and storing the number into the address corresponding to the encoding. It is not clear if this is alphabetization or not, and it was not clear how to collect the result.

## 2. THE P = ? NP PROBLEM AND TRUTH

In order to study "the basic nature of computation and not merely minor aspects of our models of computers" (Baker, 1975), the polynomial time versus non deterministic polynomial time class equivalence problem was developed by Cook(1968) and Karp(1972). The problem basically asked if the satisfiability definition of truth could be computed by a deterministic Turing machine (TM) as fast as it could be computed by a non-deterministic TM. The satisfiability conception of truth goes back to Alfred Tarski's work in the 1930s (Tarski, 1956) that defined a statement (conjunction) of basic propositions to be true if it is true under any possible assignment of truth values to the basic atomic propositions in the statement.

This problem is not only the central problem of computer science, but according to Aaronson(2005, p. 2) "is correctly seen as the deepest problem in all mathematics". Since the formulation of the P =? NP problem in the late 1960s, it has become both a mathematical problem, a scientific problem because it involves time and a philosophical problem. The "canonical" possibly easiest problem in the NP class of problems is the logical truth satisfiability problem. Following Karp, other problems in the class NP (solvable in in a polynomially bound number of steps on a non deterministic TM) are solved by mapping to the satisfiability problem in polynomial time (Karp,1972). The satisfiability problem and its characterization of what can be computed is closely related to the very essence of truth because as 18th philosopher David Hume observed, "no general proposition whose validity is subject to the test of actual experience can ever be logically certain. ... [something] substantiated in n-1 cases affords no logical guarantee that it will be substantiated in the nth case also" (Ayer,1936, p. 289).

This paper considers the epistemology of computation in the quasi-empirical sense by investigating "what is true, and not what is *hypothetically taken to be true* (for instance axioms)" (Finsler, 1996, p. 162).

## 3. Problems Solved by Computational Epistemology

Two obvious problems solved by computing are disproof of the deflationist definition of truth and disproof of the form of intuitionism that disavows the law of the excluded middle. The deflationist theory of truth (Stanford Encyclopedia, 2010) argues "to assert a statement is true is just to assert the statement itself". Computation epistemology of truth as a satisfiable assignment to all atomic elements is obviously more than merely "asserting a statement".

There are a number of forms of intuitionism. One form rejects the law of the excluded middle. It is claimed there are formulas that are neither true nor false (probably because they can not be constructed in a intuitively obvious way). Again, existence for finite formulas (possibly potentially infinite unbounded formulas also) can be tested by finding some assignment of true and false to atomic clauses that makes the formula evaluate to true. If no such assignment exists, the formula is false (Finsler, 1996, pp. 167-168). There is no question of intuitively acceptable methods here.

#### 4. Problems Unsolvable by Computational Epistemology

Although, satisfiability computable in a reasonable amount of time solves some epistemological problems, it can not deal with problems involving actual infinity. From Finsler(1996, p. 164):

*One cannot form the set of all ordinal numbers, since its definition contains an inherent contradiction [Russell's paradox]. If it were not an ordinal number, then it would still contain exactly all preceding ordinal numbers, and therefore it would have to contain itself as an element which is impossible.*

#### 5. Internal Problems of Computational Epistemology - Oracle Use

One of the first attempts to solve the  $P \stackrel{?}{=} NP$  problem tried to use an infinite counting argument from meta-mathematics (Baker, 1975). The method goes back to Cantor's diagonalization using the lack of a one-to-one mapping between real and rational numbers. The modern meta-mathematical model theory analog of diagonalization is relativization using oracles. The idea is to allow TMs to make unit time calls to an oracle. The hope was that for all oracles the class of languages recognized by  $P$  plus an oracle was strictly contained in (not one-to-one)  $NP$  with an oracle. The result was that  $P$  is in  $NP$  for some oracles but not for others. The Baker et. al. conclusion was that by "slightly altering the machine model, we can obtain differing answers" (p. 431).

Since then, much of computational complexity theory has been dedicated to relativizations because relativization proper containment immediately shows  $P \neq NP$ . Researchers who think there may be epistemological difficulties with the  $P \stackrel{?}{=} NP$  problem have criticized relativization but mostly without success (Hartmanis, 1976 & Hartmanis, 1992). Relativization pertains to computational epistemology because it removes problem specific structure from computable truth. Hartmanis(1976) shows that for models of computation that allow the use of more efficient storage access such as the MRAM model which has unit cost for multiplication,  $P = NP$  (pp. 33-46). This may show that there is some conceptual problem with the Church-Turing Thesis (definition of TMs) or even that the class  $NP$  does not really exist (it is an illusion in the Finslerian sense) because abstraction of the structural connection between satisfiability and other problems that need non deterministic computation for efficiency is incorrect.

#### 6. Physicalization of Computational Epistemology

Computational epistemology has taken a recent turn toward arguing that studying the  $P \stackrel{?}{=} NP$  problem "can yield new insights, not just about computer science but about physics as well" (Aaronson, 2005, p. 1). Deolalikar(2010) recently published a **proof** that  $P \neq NP$  except unfortunately it needed axioms from empirical theories of statistical physics.

In conclusion, I see this change in direction negatively because it attempts to convert a question from physics on the existence of quantum computers (QCs) (pp. 5-8) into formal and axiomatized computational epistemology that does not allow quasi-empirical experimentation. The argument comes full circle because the mathematicians who contributed to the development of modern physics (including Finsler whose main

area was the differential geometry of general relativity, p. vii) were skeptical of exactly the physics that QCs embody and require.

In his post WW II standard graduate level quantum mechanics text book, Leonard Schiff argues that "QM's range of applicability is limited to approximating the behavior of the atom" (Schiff, 1949, p. 267). Also, Paul Feyerabend's analysis of the theories of Niels Bohr and David Bohm (Feyerabend, 1982), show that the very properties assumed by QC builders do not exist. Bohr states (Feyerabend's italics): "*At the same time we must deny the universal validity of the superposition principle and must admit that it is but a (very useful) instrument of prediction.*" (p. 258). Also Feyerabend (David Bohm taught QM to Feyerabend) describes Bohm's view of the uncertainty principles as: "However in order to show the basic and irrefutable character of the uncertainty principle these features themselves would have to be demonstrated as basic and irrefutable." (p. 223).

## References

- Arronson, A (2005). NP-completeness problems and physical reality. *Sigact News* (vol. 36). (Also [www.scottaaronson.com/papers/npcomplete.pdf](http://www.scottaaronson.com/papers/npcomplete.pdf)).
- Ayer, A. (1936) in P. Benacerraf & H. Putnam(1964) (Eds), *Philosophy of mathematics - selected readings*, first edition (289-301), excerpt of Ayer, A. *Language, truth and logic*.
- Baker, T., Gill, J. & Solovay, R. (1975) Relativizations of the P =? NP question. *Siam J. Comput.* 11(4), 431-442.
- Cook, S.(1971) The Complexity of Theorem-proving procedures, *Proceedings of the third Annual ACM symposium on Theory of Computing*. (151-158).
- Deolalikar, V (2010) P != NP, HP Research Labs, Palo Alto, August 6, 2010, unpublished.
- Stanford Encyclopedia of Philosophy (1981) Deflationary Theory of Truth, (URL of Feb. 2011: [plato.stanford.edu/entries/truth-deflationary](http://plato.stanford.edu/entries/truth-deflationary)).
- Feyerabend, P. (1981) *Philosophical papers. Vol. 1. Realism, Rationalism & Scientific Method*, Cambridge.
- Finsler, P. (1996) in D. Booth & R. Ziegler eds.), *Finsler set theory: Platonism and Circularity*, Birkhauser.
- Hartmanis, J. & Simon, J. (1976) On the Structure of Feasible Computations, in M. Rubinoff. & M. Yovits (eds.) *Advances in Computers 14*, Academic Press, 1-43.
- Hartmanis, J. et. al. (1992) Relativization: A revisionistic retrospective. *Bulletin of the EATCS*. Vol. 47.
- Lakatos, I. (1978) *Philosophical papers. Vol. 2. Mathematics, Science and epistemology*. (ed. J. Worrall & G. Currie ), Cambridge, 24-41 (expanded version from *Proceedings of the Fourth International Congress for Logic*. ed. I. Lakatos(1967), North Holland).
- Lakatos, I. (1976) *Proofs and Refutations*. Cambridge.
- Schiff, L. (1949) *Quantum Mechanics*. First edition, McGraw Hill, New York.
- Tarski, A (1956) The Concept of Truth in Formalized Languages, *Logic, Semantics, Metamathematics*, Clarendon Press, 152-278.

## PHILOSOPHY OF THE WEB AS ARTIFACTUALIZATION

ALEXANDRE MONNIN

*Université Paris 1 Panthéon-Sorbonne (PHICO, EXeCO),*

*Institut de Recherche et d'Innovation,*

*Conservatoire National des Arts et Métiers (DICEN)*

*12, place du Panthéon*

*75231 - Paris cedex 05, FRANCE*

AND

HARRY HALPIN

*World Wide Web Consortium*

*MIT/CSAIL*

*32 Vassar St., Bldg. 32-G514*

*Cambridge, MA 02139, USA*

**Abstract.** What is the philosophical foundation of the World Wide Web? T. Berners-Lee, widely acclaimed as the inventor of the Web, has developed informal reflections over the central role of URIs (Uniform Resource Identifiers, previously Uniform Resource Locators) as a universal naming system, a central topic in philosophy since at least the pioneering works of R. Barcan Marcus. URIs (such as <http://www.example.org/>) identify anything on the Web, so the Web can be considered the space of all URIs. In a debate between Berners-Lee and P. Hayes over URIs and their capacity to uniquely 'identify' resources, Berners-Lee held that engineers decide how protocols should work and that these precisions should determine the constraints of reference and identity while Hayes held that names have their possible referents determined only as traditionally understood by logical semantics, which Hayes held engineers could not change but only had to obey. This duality can be interpreted as an opposition between a material *a priori* and a formal *a priori*. The material *a priori* of technical systems like the Web is brought about by what we call 'artifactualization', a process where concepts become 'embodied' in materiality - with lasting consequences.

## 1000-word abstract

What is the philosophical foundation of the WWW? Is it an open and distributed hypermedia system? Universal information space? How does it differ from the Internet?

While the “ecology” of the Web has known many a revolution, in contrast, its underlying architecture remains fairly stable. URIs, the HTTP protocol, resources, and languages like HTML and RDF constitute the building blocks of the Web. As the particular kind of computing embodied by the Web has displaced traditional desktop applications, the foundations of Web architecture and its relationship to wider computing needs to be clarified in order to determine both its roots, boundaries, the reasons for its success, future developments... This is especially urgent as now debate is opening over platforms and cloud computing, as how they relate to the Web.

Tim Berners-Lee, widely acclaimed as the inventor of the Web, has developed in his design notes informal reflections over the central role of URIs (Uniform Resource *Identifiers* – previously *Locators*) as a universal naming system, a central topic in philosophy since at least the pioneering works of Barcan Marcus. URIs (such as <http://www.example.org/>) identify anything on the Web so it can be considered the space of all URIs. The concrete access mechanisms of how information is transmitted via a URI is then determined by the Internet, and so the Web could be built on another architecture (such as the “Future Internet”), and likewise the Internet can also host other applications than the Web, such as peer-to-peer file-sharing.

Possible entities denoted by URIs are called *resources*. While high-order ontological debates have continuously tried to provide distinctions between endurants and perdurants (categories that mainly apply to substances), the characterization of

resources has relied on vastly different ontological principles that descend from

engineering concerns rather than claims of ontological correctness.

Drawing from the work of Vuillemin, we draw a parallel between the Web and philosophical systems. Like the former, it is concerned with traditional issues pertaining to the philosophy of language (URIs as proper names), to ontology (the link between engineering design choices in Semantic Web ontologies and philosophical ones), and metaphysics (entities of the Web as resources). Unlike philosophical systems that reflect on the constraints of the world, the Web is a world-wide embodied technical artifact that therefore creates a whole new set of constraints. We suggest that they should be understood as a material *a priori* - in the Husserlian sense - grounded in history and technology.

In a striking debate between Berners-Lee and Patrick Hayes over URIs and their capacity to uniquely ‘identify’ resources, Berners-Lee held that engineers decide how the protocol should work and that these decisions should determine the constraints of reference and identity. Hayes replied that names have their possible referents determined only as traditionally understood by formal semantics, which he held engineers could not

change but only had to obey. This duality can be interpreted as an opposition between a material and a formal *a priori*. Interestingly enough, recently Hayes is focusing on adopting principles from the Web into logical semantics itself.

The material *a priori* of technical systems like the Web is brought about by what we call “*artifactualization*”, a process where concepts become “embodied” in materiality - with lasting consequences. While such a process clearly predates the Web we can now see within a single human lifetime the increasing speed at which it takes place, and through which technical categories (and philosophical ones) are becoming increasingly dominant over “natural” and “logical” categories. At the same time, the process of having philosophical ideas take a concrete form via technology lends to them often radically new characteristics, transforming these very concepts in process. Heidegger posited a filiation between technology and metaphysics, with technology realizing the Western metaphysical project (by inscribing its categories directly into concrete matter should we add). Yet, if technology is grounded in metaphysics, it is not the result of a metaphysical movement or “destiny” (*Schicksals*) but a more mundane contingent historical process, *full of surprises and novelties*. For all these reasons, it must be acknowledged that the genealogy of the Web, as a digital information system, differs from traditional computation with regards both to the concepts at stake and our relation to them (the scientific ethos being replaced by an engineering one – something Berners-Lee dubbed “philosophical engineering”).

On the Web, the activity of standardization through bodies like the W3C arguably consists in making sense of technological evolution *post-hoc*. Nevertheless, regarding the architecture of the Web, one may argue that its standards were both the result of a process of conscious decision-making in specifying how protocols should work and the result of a constant adjustment to the reality of the technical system. Therefore, the Web can be seen as an artifact both in terms of being a designed human invention and a non-human (Latour) whose study may lead to numerous unintended discoveries, beyond its initial design.

For all these reasons, the very practice of philosophy is transformed by having to take this material *a priori* and its technical categories as seriously as “natural” or “analytic” categories from biology or natural language. Philosophers then have to deal

with technical categories that may have a lasting effect in spheres like the Web, not just

as variants from categories that can be analytically understood, but rather as concrete artifacts which can even transform the previously considered analytic categories (ironically, the main challenge to analytic judgments is no longer what Quine called “naturalization” but rather the ongoing artifactualization). While at first glance URIs can be considered just another kind of name and so inherit the characteristics and debates in philosophy over the referential status of proper names, the Web makes a difference, as URIs primarily are used to physically access information such as webpages – an aspect of naming for the most part foreign to the philosophy of language.

R. Sennett’s craftsman’s motto might be “doing is thinking”, once concepts have been artifactualized (and, as a consequence, externalized), thinking is also doing or conceiving; in the end, a matter of design.

## ONTOLOGICAL COMMITMENTS OF COMPUTER SCIENCE

MIGUEL PAGANO

*FaMAF – Univ. Nacional de Córdoba*

*Medina Allende s/n, X5000HUA Córdoba, Argentina.*

**Abstract.** We suggest that a fictionalist attitude with respect to Quine's proposal of ontological commitments is best suited for building up an ontology for computer science. In particular, we argue in favour of using theories of programming languages for identifying the relevant ontological categories.

### 1. Introduction

In this extended abstract we propose a novel reading of Quine's ontological commitments [Quine, 1980] to analyse the ontology of computer science. We argue that a fictionalist posture (see [Szabó, 2009]) can save genuine concepts of computer science from vanishing as ingenuous mathematical construction. Although we only discuss aspects related to programming languages and programs, we think that this can lead to a fruitful research programme if extended to other areas of computer science.

### 2. Programming Languages: Ontology from Semantics

Before coming to our proposal, let us briefly review critically two papers by A. Eden and R. Turner which deal with the ontology of computer science. In the first paper [Eden and Turner, 2007a] they study the ontological commitments of programming languages. They propose that semantics determine to which entities a particular programming language is committed. They apply this methodology for a simple imperative language with two kinds of semantics (based on set theory and type theory, respectively). We do agree on the use of semantics to determine some of the commitments of computer science, however it is not clear to us that programming languages have ontological commitments; instead they should be attributed to theories of programming languages (TPL). The fictionalist attitude enters here: the fact that TPL uses a certain mathematical foundation, say set-theory, does not imply that its commitments are those carried by the foundational theory; instead concepts like abstract syntax, reference, state, ordered structure given by the outcome of a certain computation are our candidates for the ontological commitments; i.e. the entities which should be used to reason about

programming languages and programs-scripts. Instead of trying to appeal to the language on which the genuine concepts are modeled, we propose to justify the commitments in terms of their epistemological value.

In the second paper [Eden and Turner, 2007b] Eden and Turner put semantics aside as the source of the commitments carried on by PL; in this article the underlying programming paradigm determines the true entities to which a programming language is committed. It can be posited that some of the aforementioned examples could be taken to be specific to some or other paradigm; but, it is not obvious to us that programming paradigms are good candidates to look for commitments. Consider, for example, what kinds of reasoning can be done by only knowing the paradigm of a PL but without any deeper theory about PL, it would be surprising that one could decide if two program-scripts compute the same or not. What is more strange to us is the attempt to attach commitments to programming languages or programs-scripts: PL are not more than the description of a set of valid programs (the so-called programs-scripts) with a notion of execution – the former usually given by a more or less abstract grammar and the latter presented by more or less formal means, ranging from a fully-formalised semantics to a mere bogus and ambiguous compiler.

We have already mentioned some ontological commitments with an epistemological basis; now we use syntax to show that TPL are the good place to look for the genuine building blocks of (part of) the ontology of computer science. In a first overview the only interesting category arising from considering syntax is that of program-scripts (cf. [Eden and Turner, 2007b]), but program-scripts alone are not enough descriptive to grasp the importance of different parts of a program-script.

For example, two occurrences of the same variable can play different rôles, say one occurrence can be a formal parameter in a procedure or function and the other an occurrence in a program calling the procedure. Just from a syntactical point of view, there should be a distinction between those two occurrences, the formal parameter is a binding occurrence, while the other occurs free occurrence. On the other hand, one could also be tempted to pay too much attention to syntax and introduce some superfluous concepts, e.g. differentiating between parsed or un-parsed program scripts or putting a two restrictive condition on what is a program-script. Since the best account of the interesting syntactical phenomena is given by abstract syntax, we should expect to get from its development [McCarthy, 1962, Fiore et al., 1999] the ontological categories corresponding to the syntactical aspects of PL.

### **3. Conclusion**

Let us conclude by commenting on how to use semantics (may be the best known area of TPL) for studying the ontology of computer science. We acknowledge that asking for a definite semantics in order to establish a new ontological category can delay the acceptance of new concepts brought by new languages lacking a proper definition and defined in terms of a compiler or interpreter. In spite of not considering the ontology as an immutable edifice, we should restrain of adding new concepts as fast as a new paradigm or PL is announced; instead we think a more parsimonious attitude should be observed and wait until a good semantic explanation is given for the newly introduced artefacts.

We do not advocate that one kind of semantics should be preferred over others, based on the status given by some foundational philosophy of mathematics to its underlying theory; Turner [Turner, 2009] seems to accept that any semantics should be accepted as a mathematical entity by a realistic mathematician. It is clear to us that the various proposed semantics could explain diverse aspects of the same language and account for several ontological categories.<sup>2</sup>

From the fictionalist posture we adopt, it is futile to try to explain in what sense the categories of a resulting ontology built up by following TPL are more relevant metaphysically than those arising from other proposals, say Eden and Turner's papers. Our proposal would correspond to what Smith [Smith, 2003] calls an "internal metaphysics" and its merits reside on how good it is for accounting the phenomena studied on computer science.

### Acknowledgements

I am grateful to Martin Diller, Pío Garcia, and Renato Cherini for encouraging me to write this abstract. My work is founded by CONICET, Argentina.

### References

- Eden, A. H. and Turner, R. (2007a). *Computation, Information, Cognition. The Nexus and the Liminal*, chapter *Towards a programming language ontology* (pp: 147–159). Cambridge Scholars Publishing.
- Eden, A. H. and Turner, R. (2007b). Problems in the ontology of computer programs. *Applied Ontology*, 2(1):1(pp: 3–36).
- Fiore, M., Plotkin, G., and Turi, D. (1999). Abstract syntax and variable binding. *Proceedings of the 14th Annual IEEE Symposium on Logic in Computer Science, LICS '99* (pp: 193–202). Washington, DC: IEEE Computer Society.
- McCarthy, J. (1962). Towards a Mathematical Science of Computation. In IFIP Congress (pp: 21–28).
- Plotkin, G. D. (2004). The origins of structural operational semantics. *Journal of Logic and Algebraic Programming*, 60-61. (pp: 3–15).
- Quine, W. V. O. (1980). *From a Logical Point of View*, chapter *On What There IS*, (pp. 1–19). Harvard University Press.
- Scott, D. S. (1970). Outline of a Mathematical Theory of Computation. Technical Report PRG-2, Oxford, England.
- Smith, B. (2003). *The Blackwell Guide to the Philosophy of Computing and Information*, chapter *Ontology*. Wiley-Blackwell.

---

2 For example, Plotkin's operational semantics leading to a better understanding of the implementation of programming languages [Plotkin, 2004] and Scott's denotational semantics [Scott, 1970] used to reason about the equivalence of programs without resorting to a particular implementation.

- Z. G. Szabó, *The Analytical Way. Proceedings of the 6th European Congress of Analytic Philosophy*, chapter The Ontological Attitude. London: College Publications, 2010.  
Available at <http://pantheon.yale.edu/~zs47/documents/Theontologicalattitude.pdf>
- Turner, R. (2009). The Meaning of Programming Languages. *American Philosophical Association Newsletter on Philosophy and Computers*, Fall-2009 (pp. 2–7).

## Semantics of Programming Languages

UWE V. RISS  
*SAP Research Karlsruhe*  
*Vincenz-Priessnitz-Str. 1*  
*76131 Karlsruhe*  
*Germany*

**Abstract.** The grounding of the semantics of programming languages is investigated. It is argued that the meaning of programming languages results from the operations that they abstract and the interpretation of these operations in terms of human activities as the final point of reference. This view opposes the interpretation of the semantics of programming languages. The latter refers to higher order abstraction as basis whereas the current view sees these semantics rooted in the actual performance realized by concrete implementations, taking a pragmatic stance.

### 1. Introduction

The central aim is to investigate the role of computers and the grounding of semantics of programming languages. Traditional approaches towards the semantics of programming languages such as operational or denotational semantics (Turner, 2007) aim at abstracting from the differences of individual implementation to find the common *meaning* behind them. Operational semantics does this by referring to abstract machines while denotational semantics refers to mathematical structures. In the following it is argued that semantics cannot be understood in such terms of higher order abstraction but, on the contrary, must be rooting in concrete operations. We can understand the mentioned approaches as objectifications of the perceived equivalence of the respective operations. However, the point of reference for semantics cannot be this objectification but the underlying concrete operations and their perceived equivalence (Saab and Riss, 2010), in analogy to the natural sciences the basis of which are experiments and not scientific laws.

### 2. Activity Theory

For this purpose we primarily regard computers as tools in human activity. The framework of this consideration is Activity Theory (Engeström, 1987) that describes the relation between persons (subjects), the objects of their activities, and the context of these activities in the schematic triangle depicted in Figure 1:

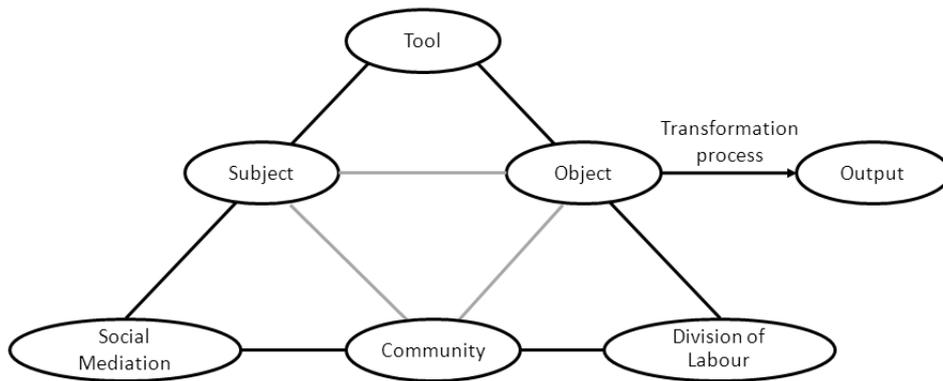


Figure 1. Activity Triangle.

The core triangle of subject (human agent), community, and object has been extended towards tools, communication (social mediation), and division of labour. All human activity is directed towards an object and aiming at a desired output. The social context includes language and communication that mediate the interaction between subject and community. Hereby communication appears as a means for activity coordination and knowledge transfer within a community and thus enables division of labour.

Understanding computers merely as tools in this system, however, is not sufficient since this neglects several specific aspects such as the separation of hardware and software. The term programming language already indicates that the concept of software is related to communication while hardware represents a traditional tool concept. Thus, programming languages serve a means of communication between the subject and the hardware representing the proper tool. This interpretation can be further supported by the objectives of artificial intelligence research to introduce intelligent agents that as equivalent to human agents regarding their intellectual capacity. Even if this goal is not reached, computers move down in the diagram from the top position (*tool*) towards a middle position where more complicated coordination and communication is required.

### 3. Fundamental Understanding of Semantics

To understand semantics of programming languages we have to go back to natural languages. These are generally used as means to coordinate the activities among collaborating human agents and to transfer knowledge; program languages are used to organise the division of labour between the human agent and the computer and to instruct the computer what to do, both at a rather elementary level. If we look at two key features of natural language, abstraction and symbolization, we also find them in programming languages. Every line of code in an ordinary computer program symbolises an abstraction of simple operations that both humans and machines can (usually) execute with equivalent results. Thus, abstraction is the key to transferability of operations from one person to another or from a person to a computer. However, abstraction must not be regarded as absolute but as a process of identification. Symbolization as the

manifestation of such identity serves as the basis of the machine's automatic processing of programs. On both sides, human agents and computers, it is the capacity to reliably interpret symbolic expressions, which ensures a repeatable execution of operations and the use of the computer as a tool.

The basis for communication via symbolized abstraction and coordination of operations is shared meaning. Here meaning of messages includes two aspects, the interpretation of messages and the expectation that others understand it in a similar way (Saab and Riss, 2011). In the case of computers it is sufficient that this expectation is one-sided, that is, from the human agent towards the machine; the computer is not supposed to have expectations. Regarding the concept of meaning we refer to a pragmatist view that understands the meaning of a message as what an agent *can do with this message* (Stegmaier, 2001). For the subject the meaning of program code is determined by the subject's knowledge of how to execute the included operations while the hardware determines the 'meaning' for the computer, that is, the computer is able to execute the program. Naturally semantics is not equated with execution – a single malfunction does not spoil the meaning of a computer program – but with execution as a repeated process of significant reliability. In the case of computers we even find a more reliable execution than what we can expect of human agents.

#### 4. Abstract Semantics

If the meaning of programming languages is not constituted by higher levels of abstraction but by concrete operations we have to clarify the role of abstract formal approaches, as they appear in operational or denotational semantics (Turner, 2007). In the same way as mathematical models abstract human activities these formal semantic model abstract operations and serves as means to support program development and testing. Formal definitions are only meaningful inasmuch as they refer to established human practice. Indeed engineers have constructed computers before researchers have applied formal semantics to programs so that formal semantics cannot be seen as the actual foundation for computer languages. Formal semantics can only support the development process but not constitute it.

The presented approach shows some links to Rapaport's idea of implementation as semantic interpretation (Rapaport, 2005). It also resembles the idea of information as sense-making of data (Saab and Riss, 2011), where programs are understood as data the meaning of which results from an interpretations process that is determined by the projected operations that refer to what the computer can do with a program.

#### References

- Engeström, Y. S. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Helsinki: Orienta-Konsultit Oy.
- Rapaport, W. J. (2005). Implementation is semantic integration: Further thoughts. *Journal of Experimental & Theoretical Artificial Intelligence*, 17(4), 385–417.
- Saab, D. J. & Riss, U. V. (2010). Logic and abstraction as capabilities of the mind. In: J. Vallverdù (Ed.) *Thinking Machines and the Philosophy of Computer Science: Concepts and Principles*, (pp 132-148). Hershey, PA: Information Science Reference.

- Saab, D. J. & Riss, U. V. (2011). Information as Ontologization. *Journal of the American Society for Information Science and Technology*. (accepted for publication).
- Stegmaier, W. (2008). *Philosophie der Orientierung*. Berlin, New York: Walter de Gruyter.
- Turner, R. (2007). Understanding programming languages. *Minds and Machines*, 17(2), 203-216.

## QUINEAN HOLISM AND THE INDETERMINACY OF COMPILATION

NATHAN SINCLAIR  
*Macquarie University*  
*Nathan.Sinclair@mq.edu.au*

### 1. Motivation

No other philosophical doctrine with even the remotest skerrick of plausibility would, if vindicated, so radically overthrow our current understanding of language, psychology and rationality as Quinean semantic holism. If individual words and sentences do not have meanings then we cannot explain communication as the transmission of ideas or judgments, nor appeal to sentence meanings as objects of putative propositional attitudes, nor explain reasoning in terms of the discernment of relationships between the meanings of premises and conclusions.

The very fact that sentence meanings are so fundamental to our current accounts of semantics, cognitive psychology, and reasoning, has meant that objections to Quinean holism which, if deployed against less radical claims, would be lightly dismissed, have been taken very seriously indeed. Most such objections appeal, broadly, to two hopes or assumptions. One the one hand it is claimed that the range of evidence proponents of Quinean holism have considered relevant to meaning and translation is too narrow, and hoped that somewhere beyond that range, perhaps in normative social practices or introspection, there is evidence to justify the attribution of determinate meanings to our words and sentences. On the other hand, it is claimed that arguments for the indeterminacy of translation must be *reductio ad absurda* because at best they show that the range of evidence considered is "unable to account for distinctions concerning the feature, meaning, which we know independently to exist" (Searle 1987).

While objections based on wishful thinking and "just knowing" would be dismissed if used to defend less well entrenched prejudices, once given any weight they have the dubious merit of stymieing further theoretical argument. No argument based upon lack of evidence is strong enough to preclude the hope of finding further evidence for such a dearly and deeply held assumption. To advance the dispute we need examples of alternative incompatible translations between theories expressed in clearly holistic languages.

Ideally, such examples of alternative translations between holistic languages would be pre-existing translations routinely employed for practical purposes, rather than philosophical inventions. Ideally also, the languages involved would be rigorously specified, with formal compositional grammars precisely delineating their well-formed formulae, and the theories would express their empirical contents so clearly and unambiguously that those contents could be mechanically determined. Even better if the

theories being translated included both small and easily understood theories, (so that we might easily see the scope and consequences of the indeterminacy of translation) and theories as large and complex as our grandest scientific theories (so we could see that the indeterminacy was not an artifact of theoretical simplicity). Better yet if each such theory could be taken as complete and self-standing, in order to ensure that the indeterminacy of translation was not the result of taking statements out of context. Astoundingly, all these desiderata are fulfilled by programming languages, compilers, and computer programs. Languages, forms of translation, and theories, so common that few of us in the developed world are ever more than arms length from tools that rely upon them for their operation.

## 2. Outline

In part one of this presentation I argue that computer programs are (readily converted into) empirical theories. Programs' empirical contents are the patterns of input and output produced by processes executing them. The under-determination of programs by their input-output is so well known and unthreatening that in many universities a high degree of similarity of program structure, even between simple programs required to produce the same output, is grounds for suspicion of plagiarism. Furthermore, programs are obviously holistic in the sense that (most) statements in computer programs do not produce any output, nor is any fragment of the output of such programs directly attributable to them. This insight allows us to make sense of the Quinean doctrine that individual sentences simply do not have meanings, and to see that the inferential/conceptual role semantics many critics (most notably Fodor and Lepore) attribute to Quine, according to which the meanings of individual sentences are determined by the theories of which they are a part, is a grotesque misinterpretation of Quinean holism.

In part two I show that compilation (and decompilation) is a form of translation by the standards Quine advocated, and then argue briefly that those standards are adequate and that compilation is translation simpliciter. I then show that the indeterminacy of compilation is well known and unthreatening to computer scientists. The only guarantee given by ISO standard compliant compilers is the preservation of input-output behaviour, and computer scientists know that independently written compilers are unlikely to produce the same machine (or high level) code given the same source code, and are unsurprised when decompilers cannot accurately reconstruct original source code. Furthermore, computer programs obviously exemplify the principles of (near) universal revisability and maintainability that philosophers have found so troubling and implausible and yet, as the practice of debugging shows, there can be good reason to revise some sentences and not others in the face of recalcitrant experience.

In part three I consider recent developments in the semantics of programming languages, whether the indeterminacy of compilation is sufficient to undermine the existence of an analytic-synthetic distinction in programming languages and argue that the translation of natural languages is less tightly determined than the translation of programming languages.

The position I advocate in this presentation is compatible with both normative and dispositional accounts of semantics. Whether the ISO standard for the C programming language is regarded as specifying dispositions possessed by C programs and compilers,

or the norms to which programs are subject once they are held to be C compilers, the compilation of C programs is (properly) indeterminate and C programs are (properly) under-determined by the input-output they are intended to produce.

In order of increasing ambitiousness, I hope people who attend this presentation will discover that Quinean holism is not a form of inferential/conceptual role semantics, computer programming languages are holistic and exemplify the controversial features of Quinean holism, compilation exemplifies indeterminate translation, and why it is plausible that translation of natural languages is even less determinate than compilation.

## References

- ISO/IEC WG14 N1256: Programming Languages – C, 2007-09-07, International Organization for Standardization, Geneva, Switzerland,  
<http://www.openstd.org/jtc1/sc22/wg14/www/standards>
- Allison, L. (1986). *A Practical Introduction to Denotational Semantics*, Cambridge University Press.
- Fodor, J. Lepore, E.( 1992). *Holism: A Shoppers Guide*. Blackwell Publishers.
- Fodor, J. (2004). Having Concepts: a Brief Refutation of the Twentieth Century. *Mind and Language* 19, no. 1 (February): 29-47.
- McDermott, M. (2009). A Science of Intention. *The Philosophical Quarterly* 59, no. 235 April: 252-273.
- Morrison, J. (2008). Just how controversial is evidential holism? *Synthese* 173, no. 3 (November 22): 335-352.
- Okasha, S. (2000). Holism about meaning and about evidence: in defence of W. V. Quine. *Erkenntnis*: 39-61.
- Quine, W. (1961). Two dogmas of empiricism. In *From a Logical Point of View*, 20-46. 2nd ed. Harvard University Press.
- Quine, W. (1964). *Word and object*. MIT press.
- Quine, W. (1977). *Ontological Relativity and Other Essays*. Columbia Univ Pr.
- Searle, J. (1987). Indeterminacy, empiricism, and the first person. *The Journal of Philosophy* 84, no. 3: 123–146..
- Winskel, G. (1993). *The Formal Semantics of Programming Languages*. MIT Press.

## IS FINDING A ‘BLACK SWAN’ POPPER, (1936) POSSIBLE IN SOFTWARE DEVELOPMENT?

LINDSAY SMITH  
University of Hertfordshire, UK  
l.l.smith@herts.ac.uk

AND

PAUL WERNICK  
University of Hertfordshire, UK  
p.d.wernick@herts.ac.uk

AND

VITO VENEZIANO  
University of Hertfordshire, UK  
v.veneziano@herts.ac.uk

### Introduction

Users’ experience of software-based technology that fails to meet their expectations is so widespread as to be a ‘commonplace’ occurrence ((Smith, 2009). However a satisfactory response from software engineering (SE) remains as elusive as ever.

In this paper we investigate the context of software engineering (SE) as a negotiation between the contradiction(s) of human subjective experience of software-based technology that relies on architecture inclusive of objectivity. For example machine programming languages that can be mathematically proven ‘Turing complete’, e.g. Church-Turing Thesis (Eden, 2007).

Consideration of the technological context of SE demands a philosophical re-evaluation of the ontological and epistemological status of SE in Computer Science (CS). We have undertaken a cross-disciplinary investigation to reposition unresolved problems in SE which potentially also opens up philosophical debate. For example if we introduce the development of software technology as a subject area for unresolved metaphysical debate. Such as the Kantian analytic/synthetic a priori dispute (Hacker, 2006). The limitations on this paper preclude explicit discussion on the ‘pros and cons’ of metaphysics for SE, or visa versa; however some basic principles echo implicitly in our discussion. For example our above comments on **objectivity**, e.g. possible for machine code and an (current?) impossibility for a priori understanding of **subjective** stakeholder software requirements. This implies Requirements Engineering (RE) practice occupies an epistemological ‘gap’ between the architectural basis of software and how it is built/used.

For our discussion one positive consequence of a cross-disciplinary approach is that novel questions can be asked. It would appear to be the case, for example that RE practitioners gaining an understanding of stakeholders’ requirements is

compatible with the Kantian epistemological classification of ‘synthetic a posteriori’ (Hacker, 2006). This raises the possibility of other epistemological explanations to questions such as why SE compares unfavourably for reliability with other engineering disciplines. For example, civil engineers can respond to unexpected circumstances in bridge construction by correcting faults, (BBC, 2000) whereas the hazards of safety critical faults in aircraft cockpit software are/cannot be addressed in an equivalent way. As Mellor, (1990) explains, the aviation industry certifies software for ‘airworthiness’ based on the ‘correctness’ of the software development process but not on the ‘correctness’ of the behaviour of software during testing.

Software development includes planning and designing artefacts but also presents SE with *predictive* type problems. For example RE identifies/selects software requirements to satisfy stakeholders’ future use of software. However RE lacks reliable or dependable tools/techniques to predict outcomes (Nuseibeh, 2000).

## **Rationale**

We are interested in why Computer science (CS) has not established scientific laws that can predict SE outcomes unlike, for example, civil engineering that relies on the established natural laws of Physics. The difference between CS and the natural science (NS) paradigm manifests in the division between observation of naturally occurring phenomenon and contending with artificially occurring phenomenon, e.g. software. Human interaction with software-based technology gives Social Science (SS) paradigm(s) (Burrell, 1979) potential ontological relevance for CS (Smith, 2010). For example both SS and CS need to observe ‘non-physical’ phenomena such as human interaction. However cross disciplinary research depends on what is optimal in a particular paradigm, for research purposes. Utilising different scientific paradigms (Hirshheim, 1989) is not straightforward. As a result we chose conservatively to employ SS to provide a dialectical analysis of contradictions in software development such as those outlined above. In particular we opposed a potential (1) ‘scientific paradigm’ of CS Eden (2007) with (2) Ethnomethodology (Ethnometh) an SS approach that challenges scientific paradigm(s) in SS (Garfinkel, 1967) and has provenance in RE research (Goguen, 1994). Our purpose is to explore the potential for obtaining leverage over limitations in understanding of software development.

## **Can a science base for software development be identified?**

For (1) to provide prediction a relevant definition of science needs to apply to CS. Reasons to doubt this possibility are raised by (2) and we consider this in the observation of artificial phenomena in software development.

The critical perspective of Ethnometh centres on the scope and meaning of science. We focus on ‘scientific method’ (SM) because this is how scientific prediction is achieved resulting in the development and acceptance of scientific theories as explanation(s) of meaning. SM is defined as a process that relies on both inductive reasoning and observable phenomena to create a hypothesis that can be tested. Prediction of events or observations is then a process of deductive reasoning relying on theory to direct hypothesis testing.

Prediction, for SE outcomes, is important and good practise in SE is implicitly 'Popperian' (Popper, 1936), e.g. software is built to be testable. However equating software testing to SM, e.g. a refutable hypothesis, is questionable (Eden,2007).

One central problem for establishing a scientific basis for software development is observation. Predictive SE, if possible, must have refutable observable phenomenon (Smith, 210). Yet any observation is via a human 'prism' hence the relevance of Ethnometh criticism of applying SM to social phenomena, e.g. human behaviour (Garfinkel, 1967). For software development human-technology interaction, e.g. input and output on a screen, is the point at which an artificial phenomenon (software) interfaces with its social environment (Smith, 2009). It is also the point where an SS paradigm that "capture(s) the basic assumptions of *coexistent* theories" Hirschheim, (1989) becomes relevant to CS.

Opposing theories in SS do not make the application of SM straightforward. However CS is currently in a unique cross disciplinary position. This is because software-based technology replaces previously existing environments/ phenomena with artificially occurring environments/phenomena. SE practice provides the means by which phenomenon such as the results of the execution of source code, are possible to observe.

SM has been applied via 'artificial' means before, such as instrument-assisted observation of otherwise unobservable phenomena. Historically scientific experimentation produced, for example, the discovery of electricity via investigating the directly unobservable magnetism (Mendelssohn, 1976). Certainly using artificial tools to 'empirically' observe naturally occurring phenomena, such as weather patterns, requires attention to both natural and artificial environments. Including SS paradigm(s) raises tantalising prospects such as the potential for SE to provide the means to observe artificial phenomenon.

## Bibliography

- BBC,[http://news.bbc.co.uk/1/hi/english/static/in\\_depth/uk/2000/millennium\\_bridge/default.stm](http://news.bbc.co.uk/1/hi/english/static/in_depth/uk/2000/millennium_bridge/default.stm), 2000, accessed 15/03/11.
- Burrell, G. "Sociological Paradigms and Organisational Analysis", Heinemann, 1979.
- Eden, A. "Three Paradigms of Computer Science", *Minds & Machines*, 17:135-167, 2007.
- Garfinkel, H. "Studies in Ethnomethodology", Prentice-Hall 1967.
- Goguen, J. "Requirements Engineering as the reconciliation of Technical and Social Issues." In *Requirements Engineering : Social and Technical Issues*. London Academic Press, 1994.
- Hirschheim, J. "Four Paradigms of Information Systems Development", *ACM*, Vol 32, Number 10, 1989
- Hacker, P.M.S. *Passing the naturalistic turn : On Quines Cul-De-Sac*", Philosophy, 2006.
- Mellor, P. "10 to the -9 and all that: The non-certification of flight-critical software.", City University London, 1990.
- Mendelssohn, K. "Science and western domination", Thames and Hudson, London, 1976.
- Nuseibeh, B., 'Requirements Engineering: A Roadmap', Proc. ICSE 2000.

Popper, K. *Conjectures and Refutations: The Growth of Scientific Knowledge*.  
Routledge, 1963.

Smith, L. “Meeting stakeholder expectations of software, or looking for the ‘Black Swan’ in software requirements”, Proc. ECAP09, 2009.

Smith, L. “Software development: Out of the black box”, Proc. ECAP10, 2010.

## **ONTOLOGY: from Philosophy to ICT and related areas.**

*Problems and Perspectives.*

SOLODOVNIK IRYNA

*PhD student of International PhD School of Humanities*

*University of Calabria*

*Pietro Bucci, 87036, Arcavacata di Rende (CS), Italy*

**Abstract.** This paper briefly highlights the development of the concept Ontology, from its philosophical roots up to its vision in the ICT field and related areas. Philosophically, Ontology is a systematic explanation of Being that describes the features of Reality. Nowadays Ontology is proliferating in organizing Knowledge of different domains managed by advanced computer tools. Ontology qualifies and relates semantic categories, dragging, however, the idea of what, since the seventeenth century, was a way to organize and classify objects in the world. Ontology maximizes the reusability and interoperability of concepts, capturing new Knowledge within the most granular levels of information representation. Ontology is subjected to a continuous process of exploration, formation of hypothesis, testing and review. Ontological thesis proposed today as true, tomorrow may be rejected in light of further discoveries and new and better arguments.

### **Philosophical background of Ontology**

Webster's Third New International Dictionary defines Ontology as "1. a Science or *study of Being*: specifically, a branch of Metaphysics relating to the Nature and relations of being; 2. a Theory concerning the *kinds of entities* and specifically the kinds of abstract entities that are to be admitted to a *language system*". Literally, the word Ontology comes from the Greek  $\delta\alpha\tau\omicron\varsigma$  (*ontos*) and  $\lambda\acute{o}\gamma\omicron\varsigma$  (*logos*), that means "speech about Being", but may also derive explicitly from  $\tau\acute{\alpha}$   $\acute{o}\nu\tau\alpha$  (entities), variously interpreted according to different philosophical points of view.

Aristotle proposed the first known *category system*, standing for a certain vision of the world in relation to what is judged to exist in practice. Heidegger conceived Ontology as a "phenomenology of the exploration" of what there "is" and in how it turns out. The ontological conceptualization, as a cohesive philosophical area, was introduced in 505-504 BC by Parmenides. He was the first to pose the argument about Being in its totality, presenting issue of the ambiguity among the conceptual level, Ontology and language. Parmenides recognized the ontological dimension as dominant able to subject to itself any other aspect of Philosophy. Over the centuries, the meaning of Ontology was changing depending on different visions and knowledge of other philosophers: Leucippus, Democritus, Plato, Aristotle, Descartes, Kant, Lorhard, Hegel,

Trendelenburg, Brentano, Stumpf, Meinong, Husserl, Heidegger, Gockel. Some of them gave more value to an absolute belief, another to empirical things, thus enriching the heritage of Philosophy with what is considered "*par excellence*" (the problem of existence in its fullest extent and universality: the relationship between particular and universal, intrinsic and extrinsic, essence and existence). "Indeed, without Ontology, Philosophy cannot be developed according to the demonstrative method. Even the art of discovery takes its principles from Ontology" (Blackwell,1963).

### Towards a new Ontology

The advent of Semantic web (Breitman,2007) aimed at multi-objective optimization of ICT environment and technological innovation in general, has coined a new vision of Ontology, so that it is considered today as "*formal, explicit specification of a shared conceptualization*" (Gruber,1995).

Ontology, intended as a first-order axiomatic theory expressed by a descriptive logic, is fundamental to design advanced Knowledge Based software systems (Guarino,1998; Eden,Turner,2005). It is of great interest to combine lexical resources, such as Thesaurus (Broughton, 2006) with the world knowledge provided by Ontologies in order to improve deductive reasoning with natural language, as well as enhance automatic classification (e.g. in Ontology-based Cataloging systems), problem solving techniques, interoperability among different computer systems, cross-cultural and intercultural communication in CMC (Ess, Sudweeks,2005) etc. Since Ontology is the basis of web intelligence, it is also widely used in e-commerce, on-line marketing, business management etc.

In Fig.1 we can observe philosophical reflection in the field of computer science and information technology (Floridi,2002; Colburn,2003; Gruber,2009). Here Thought (which is regulatory/normative to Reality) through Language (which defines the existing categories reflecting Thought and Reality) is connected with Ontology and Epistemology, representing the descriptive and prescriptive approaches. Ontology refers to objective validity (Husserl,1992) of terminology waiting to be discovered by domain knowledge experts and Epistemic (providing model reasoning in class-based representation formalisms through description logics).

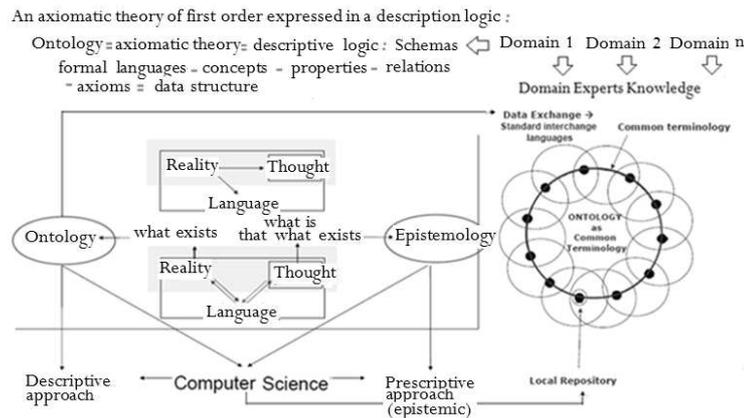


Figure 1. The ontological and epistemological turn in Computer Science

Automated reasoning and Ontology manipulation in description logics allow to present and emulate the human logic-based knowledge of entities in different domains, managing simultaneously dissimilar types of objects (concrete and abstract, independent and dependent) and their ties (relations, dependencies and predications).

Creation of *single knowledge sharing paradigm* is not easy nor immediate task, considering also non-trivial technological obstacles (consistency and validity of Ontologies vs. time and evolution of information technology). It remains an appealing challenge to set up new scientific environments in which philosophers and other scholars can meet to discuss and develop strategies to classify, organize and implement qualitative conceptual domains, and even more those represented by different semantic systems tied with language differences.

## References

- Breitman, K.; Casanova, M.; Truszkowski, W. (2007). *Semantic Web: Concepts, Technologies and Application*. NASA Systems and Software Engineering Series. 1 ed. London, Springer Verlag.
- Broughton V. (2006). *Essential thesaurus construction*, London, Facet.
- Colburn, T.R. (2003). *Philosophy and Computer Science*, Armonk, Sharpe.
- Eden, A.H. & Turner, R. (2005). Towards an Ontology of software design: The Intension/Locality Hypothesis, *3rd European Conf. Computing And Philosophy ECAP*, 2-4 Jun, Västerås, Sweden.
- Ess, C. & Sudweeks, F. (2005). Culture and computer-mediated communication: Toward new understandings. *Journal of Computer-Mediated Communication*, 11(1).
- Gruber, T. (2009). Ontology. In: *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag.
- Guarino, N. (1998), Formal Ontology and Information Systems. In: N. Guarino (Eds), *Formal Ontology in Information Systems. Proceedings of FOIS 1998*, Trento, Italy, 6-8 Jun, Amsterdam, IOS Press.
- Husserl, E. (1929). *Formal and transcendental logic*. English translation: The Hague, Martinus Nijhoff (1969).
- Smith, B. (2003b). Ontology. In: L. Floridi (Eds), *Blackwell Guide to the Philosophy of Computing and Information*, Oxford: Blackwell.
- Wolff, C. (1728). *Preliminary discourse on philosophy in general*. Translated, with an Introduction and notes, by Richard J. Blackwell (1963) Indianapolis, The Bobbs-Merrill Company.

## THE EVOLUTION OF SOFTWARE AGENTS AS DIGITAL OBJECTS

SABINE THÜRMELE

*Graduate Center of the TUM School of Education  
Technical University of Munich, Munich, Germany*

**Abstract.** The evolution of software agents as digital objects from simple interface agents to full blown interaction partners is depicted. An outline of concretization process in agent-oriented programming is given contributing to the research into the ontology of computer programs.

### Extended Abstract

The focus of this paper is on the evolution of software agents as digital i.e. computational objects. It can be shown that a new type of interplay between human beings, „computational objects“ and the physical environment is in process of emerging. Turkle’s insight (2006) into the nascent robotics culture is equally valid for software agents: „computational objects simply do things *for* us, but they do things to *us* as people, to our ways of seeing ourselves and others. Increasingly, technology puts itself into a position to do things *with* us” (p.1).

The starting point of this evolution was constituted by interface agents providing assistance for the user or acting on his or her behalf. As envisioned by (Laurel, 1991) and (Maes, 1994) they evolved into increasingly autonomous agents. In game worlds they were first seen in one person offline video games. Interacting pure software agents and avatars became prevalent in MMORPGs (massively multiplayer online role-playing games) as World of Warcraft<sup>®</sup>. As interworking collaborative software agents embedded in nets of devices they provide support for smart grids (Mainzer, 2010) or for other variants of the “Internet of things” (Mattern/Langheinrich, 2008). Last but not least they are used to coordinate emergency response services in disaster management systems (Jennings, 2010).

Already in 1992 Solum posed the question in the North Carolina Law Review whether virtual agents may be the basis for persons in the legal sense of the law (Solum, 1992). Today virtual agents are commonly deployed in online auctions or eNegociations (Woolridge, 2009). Thus software agents have been promoted from assistants to virtual interaction partners. The socio-technical fabric of our world has been augmented by these collaborative systems.

The goal of the agent-oriented programming paradigm is the adequate and intuitive modeling and implementation of complex interactions and relationships. Software agents were introduced by Hewitt's Actor Model (Hewitt et al., 1973). Today a whole variety

of definitions for software agents exist but all of them include mechanisms to support persistence, autonomy, interactivity and flexibility. Bionic approaches, as swarm intelligence, or societal models are adapted to implement collaborative approaches to distributed problem solving.

They are on the one hand part of the tool kit used in computational sciences using computer-based simulations as a link between theory and experiment. As such they are similar to numerical simulation but using different conceptual and software models.

On the other hand they provide a basis for agency in virtual worlds offering novel experiences. They provoke us to ask how this technological progress will affect our interpersonal relationships (Turkle, 2011).

The starting point of any software agent-based approach is a bionic or societal metaphor for distributed problem solving. The resulting computer science concept is specified as a computer program modeling the interacting software agents. At compile-time the high level program is transformed in a machine-executable computer program to be run in a distributed environment. During runtime any (instance of) a software agent may be perceived as a distinct thread or process. This concretization process conforms to the program abstraction taxonomy introduced in (Eden and Turner, 2007).

From an ontological perspective it can be stated that the underlying computer science concepts are abstract objects that can be concretized by computer programs conforming to an agent oriented programming paradigm. The computer programs are abstract objects that can be concretized by adequate computational objects conforming to a (different) programming paradigm or by concrete physical objects. Different concretizations may exist for one computer program. It should be noted that the identical agent-oriented program may be first tested in a simulated environment and then employed in a realtime environment.

Similar to (Reicher-Marek 2009) three basic relations between computer programs and other objects may be distinguished: the above outlined the concretization relation, the notation relation (between the abstract object and the (textual or graphical) specification), the environmental relation (between the abstract object and its potential runtime environments) and the instantiation-at-runtime relation coupling the abstract object to its dynamic instantiations. In my view any non trivial identity notion for computer programs has to take these relationships into account.

## References

- Eden, A. H. & Turner, R. (2007). Problems in the ontology of computer programs. In: *Applied Ontology*, Vol. 2, No. 1, pp. 13–36. Amsterdam: IOS Press.
- Hewitt, C. & Bishop, P. & Steiger, R. (1973). A universal modular actor formalism for Artificial Intelligence. In: *International Joint Conferences on Artificial Intelligence*, 235-245.
- Jennings, N. (2010) *ALADDIN End of Project Report*, [www.aladdinproject.org](http://www.aladdinproject.org), Cited 25 April 2011.
- Laurel, B. (1991) *Computers as theatre*, New York: Addison-Wesley.
- Maes, P. (1994) Agents that reduce work and information overload. In: *Communications of the ACM* 37 (7), 30-40.
- Mainzer, K. (2010) *Leben als Maschine? Von der Systembiologie zur Robotik und Künstlichen Intelligenz*, Paderborn: Mentis Verlag

- Mattern, F. & Langheinrich, M. (2008) Eingebettete, vernetzte und autonom handelnde Computersysteme: Szenarien und Visionen. In A. Kündig and D. Bütschi (Eds), *Die Verselbständigung des Computers*, pp. 55-75. Zürich: vdf Verlag
- Reicher-Marek, M. (2009) What is the object in which copyright can subsist? An ontological analysis. In: E. Ortland, Eberhard and R. Schmücker (Eds) *Copyright & Art. Aesthetical, legal, ontological and political issues*. Baden-Baden: Nomos, 2009. (to appear)
- Solum, L. (1992) Legal personhood for artificial intelligences. *North Carolina Law Review*, 2, 1231-1283.
- Turkle, S. (2006) *A nascent robotics culture: new complicities for companionship*. Paper presented at the 21st National Conference on Artificial Intelligence. Boston, July, 2006
- Turkle, S. (2011) *Alone Together, Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.
- Woolridge, M. (2009) *An Introduction to MultiAgent Systems* ( 2<sup>nd</sup> ed). New York: John Wiley & Sons.

## **MACHINES and COMPUTATIONS**

RAYMOND TURNER

*Department of Computer Science and Electronical Engineering*

*University of Essex*

*Wivenhoe Park, Colchester*

*Essex CO4 3SQ*

*UK*

### **Abstract**

How may abstract and physical machines be related? What is the difference between considering an abstract machine as:

1. A theory of a physical one
2. A functional description of one
3. A specification of one?

Do these distinctions throw any light on the nature of physical computation and the arguments of Putnam and Pancomputationalism?

# **Track II: Philosophy of Information and Cognition**

## ON THE LEVEL OF CREATIVITY

### *Ponderings on the Nature of Kantian Categories, Creativity and Copyrights*

ALEXANDER FUNCKE

*Centre for Study of Evolutionary Culture at Stockholm University  
106 91 Stockholm*

**Abstract.** The relation between data and information is considered in analogy with Kantian transcendental aesthetics in order to create a formal concept and ordinal relation of “creativity”. Implications are discussed for Kantian categories, creativity and copyrights.

### 1. Background & Aims

Creativity is a popular concept for controversy in many disciplines. This paper does not necessarily contain the deepest insights, but it provides perspectives that might be useful while considering creativity and thereby copyrights cognition and maybe even consciousness.

### 2. Transcendental aesthetics

In order to formulate the ideas this paper uses an analogy to Kant's transcendental aesthetics, i.e. the process where noumenon is transcended via categories to a phenomenon is contrasted to a process where data is rendered via a context/algorithm to information.

The analogy lends itself to be considered as an extension rather than an analogy of the transcendental aesthetics too. That is Kant's transcendental aesthetics may be reinterpreted as “actual” transcendence in terms of data and information. It opens up for a multiple layer interpretation, and thereby also for questions like, if we may consider a hearing aid, or other more intricate cyborg technologies as just another category in the Kantian sense.<sup>3</sup>

---

<sup>3</sup> This may also have consequences for copyrights. Arguably, copyrights ought not to be applicable to data in itself, but only to information. Now, if a blind person somehow manages to copy a protected image, then it couldn't be considered an infringement, as he lack the categories to render the information that could have

### 3. Potentiality/actuality

The dichotomy of potentiality and actuality has been part of the philosophical discussion at least since Aristotle's book *Theta*. The transcendental aesthetics analogy may be considered as a model for consider data in its actual form and its potential one relative to a given interpreter.

The interpreter in the model consist of two components, a passive *presentation* that takes *formatted data* as input and outputs *information*, and an active *algorithm* that takes *raw data* as input and outputs *formatted data*. Where the latter component may have *potential*.

An algorithm is considered to have *potential* if it manipulates the *raw data* in a way that cannot be described as a simple transformation or crop, but which also adds “extra relevant information” relative to a given *presentation*.

To formalise this potentiality, or creative quality if you will, let  $X$  and  $Y$  be sets of data, and let  $f, g \in F_{X,Y} = \{f : X \rightarrow Y\}$  be two algorithms that transforms *raw data* to *formatted data*.

Further, let  $F_{X,Y}^N \subseteq F_{X,Y}$  be the subset of algorithms that lack potential, and  $Y' \subseteq Y$  be the set of all *formatted data* that renders *information* for a given *presentation*.

Now, define two functions,  $H : X \rightarrow \mathfrak{R}$ , which maps any data to its entropy and  $H_m : Y' \rightarrow \mathfrak{R}$ , defined as

$$H_m(y) = \min_{f \in F_{X,Y}^N} H(f^{-1}(y)), \quad (1)$$

which maps any information entity to its minimal entropy representation given a *presentation*.

The inverse of  $f$  may actually not be unique, but with a small violation of notation, we define  $f^{-1}(y) = \operatorname{argmin}_{x \in \{x: f(x)=y\}} H(x)$ , that is to be the minimal entropy  $x$  that maps to  $y$ .

Finally, define the “additional map”  $A : F_{X,Y} \times Y' \rightarrow \mathfrak{R}$  such that

$$A(f, y) = H_m(y) - H(f^{-1}(y)), \quad (2)$$

---

been protected by a copyright for someone with visual categories. Nor should his original visual works ever be copyrightable for its visual qualities.

which gives a number for the level of potential the algorithm  $f$  has to generate information entity  $y$ .<sup>4</sup>

An algorithm  $f$  is considered *strictly potential* relative to a representation and a subset of the informative entities  $S \subseteq Y'$  if all its elements  $y \in Y''$  are represented more economically than in the minimal non-potential case, that is,

$$\forall y \in S, A(y) > 0, \quad (3)$$

An algorithm is considered *potential* (in the non-strict sense) for a subset  $S$  if a non-empty subset  $S' \subseteq S$  is strictly potential and for no  $y \in S, A(y) < 0$ .

#### 4. Creativity as an ordinal relation

There are various degrees of potentiality, not only should algorithm potentiality be compared with respect to the amount of relevant information quantified by the “additional map”, it should also take an interest in the relative ease to compute  $f(x) \in Y'$ .

Ignoring the complexity of computation would be like ignoring the difference between factorising the product of two huge prime and summing them.

Another example that highlights the need to include complexity is simulations of non-linear dynamical systems, such as models of meteorological or financial system. It is unfeasible to do analytical reasoning about the behaviour of such systems, and it takes a lot of computation to unfold the behaviour through simulation, even though all data and the algorithms are in place.<sup>5</sup>

There are multiple reasonable ways to define an ordinal relation between two algorithms that take these things into account,  $f, g \in F_{X,Y} = \{f : X \rightarrow Y\}$ , but the transitive, reflexive and identity preserving variant suggested here is the following,

$$f > g \Leftrightarrow O(f) > O(g) \vee (O(f) = O(g) \wedge A(f) > A(g)), \quad (4)$$

where  $O(f)$  is the computational complexity of  $f$ .

---

<sup>4</sup> Note that this means that a verbose representation  $x \in X$  of an informative entity could be classified as a non-potential, even it seem to have all the necessary properties. One could add a proxy-step to solve this, by mapping  $f$  to  $f_h$ , where  $f_h$  is the equivalence class (in the obvious sense) version of  $f$ .

<sup>5</sup> It is really just a way of stating that the tragedy of deduction will not help.

## 5. Conclusion

The concepts presented and to some extent explored in the longer version of this paper, gives a formal interpretation of the notoriously hard to pin down idea of creativity. The ordinal relation “level of creativity” lends itself to demarcate when a set of algorithms may create information that is creative enough to be regarded as copyrightable, or maybe even what is the minimal level of creativity for a cognitive or conscious algorithm?

From the analogy to transcendence there spring other implications hinted at in the footnotes: Cyborg technology, such as hearing aids may be considered as a multi-level version transcendence. Which aids ones intuition while pondering about copyrights - whether one likes Kant or not.

## References

- Dennett, D. C. (1996). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon & Schuster.
- Floridi, L. (2004). Open problems in the philosophy of information. *Metaphilosophy*, 35(4):554–582.
- Floridi, L. (2008). *Philosophy of Computing and Information: 5 Questions*. Automatic Press/VIP, Copenhagen, Denmark, Denmark.
- Floridi, L. (2009). Philosophical conceptions of information. *Lecture Notes in Computer Science*, 5363:13–53.
- Kant, I. (2003). *Critique of Pure Reason*. Courier Dover Publications.
- Koepsell, D. R. (2000), *The Ontology of Cyberspace: Law, Philosophy, and the Future of Intellectual Property*, Open Court Publishing Co., Peru, IL, USA.
- Mandelbrot, B. (1967). How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension. *Science*, 156(3775):636–638.
- Mitchell, T. (1997). *Machine Learning*. McGraw-Hill Education (ISE Editions), 1st edition.
- Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83(October):435–50.

## THE FOURTH REVOLUTION AND SEMANTIC INFORMATION

VALERIA GIARDINO

*Institut Jean Nicod (CNRS-EHESS-ENS), Paris*

*Valeria.Giardino@ens.fr*

**Abstract.** In his work, Floridi introduces several notions to describe our relationship with information and technology. Indeed, according to him, in recent times, humanity has experienced a fourth revolution, the Information revolution, which, starting from the work of Alan Turing, has deeply affected our understanding of ourselves as agents. Our generation is still a generation of “e-migrants”, but our children will be born in the infosphere and will recognize themselves from their birth as inforgs. I will focus on the notions of infosphere and inforgs, and more generally on the notion of information Floridi makes use of. According to Floridi, in re-ontologizing ourselves as inforgs, we recognize how significantly but not dramatically different we are from smart, engineered artifacts, since we have, as they have, an informational nature. Nevertheless, if one focuses on semantic information, which requires meaning and understanding, then there is still a dramatic difference between ourselves and our artifacts to be acknowledged: we are the only agents who spontaneously reason semantically. First, I will present the four revolutions Floridi talks about, and claim that there are other revolutions in the history of human culture that should be considered in the perspective of discussing the reshaping of our new environment and of our new selves in the infosphere. Secondly, I will discuss an ambiguity in Floridi’s use of the term information and propose to consider his fourth revolution as the Second Information revolution. To solve this ambiguity, I will distinguish between information and semantic information, which implies meaning and understanding. Finally, I will present some questions that emerge once we consider humans’ cognitive capacities to access meaning on the background of the new context, the infosphere.

### 1. Introduction: we are inforgs in an infosphere

Floridi has suggested that in recent years we have gone, together with our environment, through a process of re-ontologization that has changed forever our way of seeing the world and ourselves. If the challenge of philosophy today is to analyse how this revolution has changed our understanding of the world and of ourselves, my challenge in this talk will be to claim that some of Floridi’s suggestions should be partly revised and further discussed.

First, I will present the four revolutions Floridi talks about, and claim that there are other revolutions in the history of human culture that should be considered. Secondly, I

will discuss an ambiguity in Floridi's use of the term information and propose to consider his fourth revolution as the Second Information revolution. To solve this ambiguity, I will distinguish between information and semantic information, which implies meaning and understanding.

## **2. One, two, three... many revolutions: human culture**

Though I am in general sympathetic with Floridi's rational reconstruction of the four revolutions, I want to argue that in the course of human cultural evolution, it is possible to individuate other crucial steps in the transformation of our ontology.

It is unquestionable that the appearance of cognitive artefacts has played a major role in the shaping of our world and of us as cognitive agents. We might assume an evolutionary perspective and consider first the moment in which human beings began to communicate by means of a language, and then the moment they invented writing, and thus began not only to produce words but to share them in a public format that could be inspected by others and stored in archives. Both these steps were crucial in the evolution of human cognition, since they revolutionized human beings' access to meaning: new channels became available to communicate and to make sense of the world around us and of ourselves.

My approach is in line with the idea that cognition is 'distributed': as Hutchins (1995a; 1995b) explains, cognitive events are not encompassed by the skin or skull of an individual. There exist interesting kinds of distribution of cognitive processes: we must consider them if we want to understand human cognition. Human beings, despite the limitations of the cognitive systems with which we know that they are born (Kinzler and Spelke (2007); Spelke (2004)), were able to develop new practices and new cognitive strategies to augment the powers of their minds, showing an extraordinary capacity in creating tools that would help them in the processes of both describing the world around them and acting upon it. Some of these tools had an intrinsically cognitive function.

As a consequence, a more faithful reconstruction of our cultural evolution would rather show how the history of our cognition has been deeply influenced by the fact that from the very beginning we engaged ourselves in symbolic activities, and that these activities have become, in a long historical and cultural process of creation and selection, more and more complex. This was indeed a revolution in the ontology of information in the billions of years of the evolutionary process, from the time when living processes became encoded in DNA sequences: "because this novel form of information transmission was partially decoupled from genetic transmission, it sent our lineage of apes down a novel evolutionary path - a path that has continued to diverge from all other species ever since" (p. 45).

## **3. Cognition and semantic information**

In the DNA double helix, as well as in Turing machines, information is conceived as a code, a string, and it does not have anything to do with meaning or understanding. By contrast, semantic information requires meaning and understanding. Floridi claims that, by re-ontologizing ourselves as inforgs, we recognized how significantly but not

dramatically different we are from smart, engineered artifacts, since we have, as they have, an informational nature. But of what kind of information is Floridi talking about when he refers to 'informational nature' in the two cases?

I will consider Bruner (1990)'s point of view on what he defined the Cognitive revolution, taking place in the 1950s. According to Bruner's reconstruction, the aim of that revolution at the beginning was to discover and describe formally the meanings that human beings were able to create out of their encounters with the world. The objective in the long run was to propose hypotheses about which meaning-making processes were implicated in humans' cognitive activity. Bruner's hope was that such a revolution, as it was conceived at its origins, would have brought psychology to collaborate with its sister interpretative disciplines such as the humanities and the social sciences. It is only a collaboration of this kind that can allow the investigation of such a complex phenomenon as meaning-making. But the happy ever after did not work out. In fact, the emphasis began shifting from the construction of meaning to the processing of information, which are profoundly different matters.

The notion of computation was introduced and computability became 'the' good theoretical model; this brought far from the original question - the revolutionary one - which was about the conditions of our meaning-making activity, the answer of which would have explained our semantic power. For this reason, the Cognitive revolution "has been technicalized in such a manner that even undermines that original impulse" (p.1): it has become the (uninteresting) Information revolution. Meaning is thus different from information because it does not come before the message, but it is through the message itself and the fact that this message is shared that it originates. In fact, public meanings are the result of a negotiation.

#### **4. Conclusions**

To sum up, in my talk, I will try to show that a particularly interesting aspect to discuss in this framework is the role of semantic information, which is the expression of a symbolic activity that up to now has been shown to be specifically human. Knowledge is situated-distributed, and this not only because it has a cultural nature, but also and most of all because our knowledge acquisition has a cultural nature. Moreover, knowledge has also a social nature, because it gets socially constructed (Berger and Luckmann (1966)). Human beings are semantic engines, and they engage themselves in meaning-making and meaning-negotiating. For this reason, meaning is flexible: as Bruner says, we show a 'dazzling', intellectual capacity for envisioning alternatives.

Will one day a fifth revolution come that will take away from us also this ultimate illusion? That day, will our own technology bring about intentional and semantically powerful machines? At the moment, we do not know. The task of philosophy of information is to provide the appropriate framework that would allow us to make useful predictions in order to prepare the future generations and ourselves.

## Acknowledgements

I thank the *Public Representations* group at the *Institut Jean Nicod* for all our useful discussions on similar topics, and in particular Elena Pasquinelli and Giuseppe A. Veltri who read a preliminary version of this article. The research was supported by the European Community's Seventh Framework Program ([FP7/2007- 2013]) under a Marie Curie Intra-European Fellowship for Career Development, contract number no. 220686—DBR (Diagram-based Reasoning).

## References

- Berger, P. L. & T. Luckmann (1966), *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*, Garden City, NY: Anchor Books.
- Bruner, J. (1990). *Acts of Meaning*. Cambridge, Mass. and London: Harvard University Press.
- Deacon, T. W. (1997). *The Symbolic Species*. New York and London: W. V. Norton Company.
- Dror, I.E. & Harnad, S. (Eds.) (2008). *Cognition Distributed: How Cognitive Technology Extends Our Minds*. Amsterdam: John Benjamins.
- Floridi, L. (2002). Information Ethics: An Environmental Approach to the Digital Divide. *Philosophy in the Contemporary World*, 9(1), 39-45.
- (2007). A look into the future impact of ICT on our lives. *The Information Society*, 23(1), 59-64. An abridged and modified version was published in TidBITS.
- (2009). The Semantic Web vs. Web 2.0: a Philosophical Assessment. *Episteme*, 6, 25-37.
- Hutchins, E. (1995a). *Cognition in the Wild*. MIT Press.
- (1995b). How a cockpit remembers its speeds, *Cognitive Science*, 19, 265-288.
- Kinzler, K. D., & Spelke, E. S. (2007). Core systems in human cognition. *Progress in Brain Research*, 164, 257-264.
- Spelke, E. S. (2004). Core knowledge. In: N. Kanwisher & J. Duncan (Eds), *Attention and Performance: Functional neuroimaging of visual cognition* (Vol 20, pp. 29-56). Oxford: Oxford University Press.

## EPISTEMOLOGICAL AND PHENOMENOLOGICAL ISSUES IN THE USE OF BRAIN-COMPUTER INTERFACES

RICHARD HEERSMINK

*PhD Candidate*

*Macquarie Centre for Cognitive Science*

*Macquarie University, Sydney, Australia.*

*Email: richard.heersmink@gmail.com*

**Abstract.** Brain-computer interfaces (BCIs) are an emerging and converging technology that translates the brain activity of its user into command signals for external devices, ranging from motorized wheelchairs, robotic hands, environmental control systems, and computer applications. In this paper I functionally decompose BCI systems and categorize BCI applications with similar functional properties into three categories, those with (1) motor, (2) virtual, and (3) linguistic applications. I then analyse the relationship between these distinct BCI applications and their users from an epistemological and phenomenological perspective. Specifically, I analyse functional properties of BCIs in relation to the abilities (particularly motor behavior and communication) of their human users, asking how they may or may not extend these abilities. This includes a phenomenological analysis of whether BCIs are experienced as transparent extensions. Contrary to some recent philosophical claims, I conclude that, although BCIs have the potential to become bodily as well as cognitive extensions for skilled users, at this stage they are not. And while the electrodes and signal processor may to a variable degree be transparent and incorporated, the BCI system as a whole is not. Contemporary BCIs are difficult to use. Most systems only work in highly controlled laboratory settings, require a high amount of training and concentration, have very limited control options, have low and variable information transfer rates, and effector motions are often slow, clumsy and sometimes unsuccessful. These drawbacks considerably limit their possibilities for transparency and incorporation into either the body schema or cognitive system which is essential for bodily and cognitive extension. Current BCIs can therefore only be seen as a weak or metaphorical extension of the human central nervous system. To increase their potential for cognitive extension, I give suggestions for improving the interface design of what I refer to as linguistic applications.

## 1. Introduction: Brain-Computer interfaces

BCIs are an emerging and converging technology that translates the brain activity of its user into command signals for external devices. Invasive or non-invasive electrode arrays detect an intentional change in neural activity, which is translated by a signal processor into command signals for applications such as wheelchairs, robotic hands, environmental control systems, and computer applications. In essence, BCI technology establishes a direct one-way communication pathway between the human brain and an external device, and can to some extent translate human intentions into technological actions without having to use the body's neuromuscular system. However, contemporary BCIs are difficult to use, the technology is still in its infancy and has barely passed the "proof of concept" stage. Most systems only work in highly controlled laboratory settings, require a high amount of training and concentration, have very limited control options, have low and variable information transfer rates, and effector motions are often slow, clumsy and sometimes unsuccessful.

## 2. Goals, Method and Structure

### 2.1. A TYPOLOGY OF BCIS

In this paper I explore the relationship between BCI technology and their human users from an epistemological and phenomenological perspective. My analysis has five parts. First, I present a preliminary conceptual analysis of BCIs in which I functionally decompose BCI systems and categorize BCI applications with similar functional properties (Vermaas & Garbacz, 2009). Based on this preliminary analysis, I distinguish between three categories: (1) *motor applications*, which restore motor functions for disabled subjects such as motorized wheelchairs or robotic hands; (2) *linguistic applications*, which allow a disabled subject to select characters on a screen, thereby restoring communicative abilities; and (3) *virtual applications*, which allow a subject to control elements (e.g. avatars) in a virtual environment.

### 2.2. THE CURRENT DEBATE ON BCIS

Second, I briefly outline the current philosophical debate on BCIs. It has been claimed that a BCI-controlled robotic arm is a bodily extension fully integrated into the body schema of a macaque, thereby constituting a "new systemic whole" (Clark, 2007). It has also been claimed that functionally integrated BCIs are cognitive extensions, i.e., they extend cognitive processes of their users into the material environment (Fenton and Alpert, 2008; Kyselo, 2011). These philosophical claims are evaluated later on in this paper.

### 2.3. HUMAN-TECHNOLOGY RELATIONS

Third, I introduce some key concepts for better understanding human-technology relations. These key concepts are "body schema", "incorporation", "transparency" and "extended cognition". A body schema is a non-conscious neural representation of the body's position and its capabilities for action. We are able to incorporate artifacts such

as hammers, screwdrivers, pencils, walking canes, cars, glasses, and hearing aids into our body schema, thereby enlarging our body schema (Brey, 2000). These artifacts are *embodied* and are not experienced as objects in the environment but as part of the human motor or perceptual system. When using embodied artifacts to act on the world such as hammers, pencils, and screwdrivers, a subject doesn't first want an action on the artifact and then on the world. Rather, a subject merely wants an action on the world through the artifact and doesn't consciously experience the artifact when doing so. The perceptual focal point is thus at the artifact-environment interface, rather than at the agent-artifact interface (Clark, 2007). In this sense, embodied artifacts are transparent (Ihde, 1990).

Cognitive artifacts such as calculators, computers, and navigation systems, can under certain conditions be incorporated in the human cognitive system in such a way that they can best be seen as literally part of that system. These devices, then, perform functions that are *complementary* to the human brain (Sutton, 2010). There is, furthermore, a two-way interaction when using such devices, and both the brain and the cognitive artifact have a causal role in the overall process, thereby forming a "coupled system". In such coupled systems, the cognitive process is *distributed* across brain and artifact, and the artifact is seen as co-constitutive of the extended cognitive system. Remove the technological element from the equation and the overall system will drop in behavioural and cognitive competence. So there is a strong symbiosis and reciprocity in coupled systems. Moreover, what is essential when extending cognition is a high degree of trust in, reliance on, and accessibility of the cognitive artifact (Clark & Chalmers, 1998).

#### 2.4. HUMAN-BCI RELATIONS

Fourth, I explore the relationship between motor, linguistic, and virtual applications and their human users in the light of the concepts just introduced. I analyse whether BCIs are incorporated into the body schema or cognitive system of their users, and analyse whether they are experienced as transparent extensions of the human body or cognitive system. I demonstrate that, although BCIs have the potential to become bodily as well as cognitive extensions for skilled users, at this stage they are not. And while the electrodes and signal processor may to a variable degree be transparent and incorporated, the BCI system as a whole is not. Contemporary BCIs are difficult to use. Most systems only work in highly controlled laboratory settings, require a high amount of training and concentration, have very limited control options, have low and variable information transfer rates, and effector motions are often slow, clumsy and sometimes unsuccessful. These drawbacks considerably limit their possibilities for transparency and incorporation into either the body schema or cognitive system which is essential for bodily and cognitive extension.

#### 2.5. DISTRIBUTED COGNITION FOR IMPROVING BCIS

And fifth, I give suggestions to increase the potential for cognitive extension of linguistic applications. To do so, I draw from concepts of the distributed cognition framework. Jim Hollan, Ed Hutchins and David Kirsh (2000) argue that the nature of external representations is essential when effectively distributing cognition. Their notion of "history enriched digital objects" implies that often selected letters should be presented larger or brighter on the screen. Their notion of "zoomable multiscale interfaces" implies

that for someone who is selecting letters on a screen, it might be more effective if the letter the person wants to select becomes larger when the cursor moves towards it. And their notion of “intelligent use of space” implies that for people who are not used to the QWERTY-style, it might be logical to present the most often selected letters in the middle and letters that are selected less often in the periphery of the screen.

## References

- Brey, P. (2000b). Technology and Embodiment in Ihde and Merleau-Ponty. In: C. Mitcham (Ed.), *Metaphysics, Epistemology, and Technology. Research in Philosophy of Technology Vol 19*. London: Elsevier/JAI Press
- Clark, A. (2007). Re-Inventing Ourselves: The Plasticity of Embodiment, Sensing and Mind. *Journal of Medicine and Philosophy*, 32(3), 263-282.
- Clark, A. & Chalmers, D. (1998). The Extended Mind. *Analysis*, 58, 10-23.
- Fenton, A. & Alpert, S. (2008). Extending Our View on Using BCIs for Locked-in Syndrome. *Neuroethics*, (1)2, 119-132.
- Hollan, J. & Hutchins, E. & Kirsh, D. (2000). Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research. *Transactions on Computer-Human Interaction*, 7(2), 174-196.
- Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth*. Indiana University Press.
- Kyselo, M. 2011. Locked-in Syndrome and BCI: Towards an Enactive Approach to the Self. *Neuroethics*. doi:10.1007/s12152-011-9104-x.
- Sutton, J. (2010). Exograms and Interdisciplinarity. In R. Menary (Ed.), *The Extended Mind*. MIT Press.
- Vermaas, P. E. and Garbacz, P. (2009). Functional decomposition and mereology in engineering. In: A. Meijers (Ed.), *Handbook of the Philosophy of Technology and Engineering Sciences*. Elsevier: Amsterdam.

## AN INFORMATION-THEORETIC MODEL OF CHUNKING

DANIEL HEWLETT  
*University of Arizona*  
*Tucson, AZ 85745, USA*

AND

Paul Cohen  
*University of Arizona*  
*Tucson, AZ 85745, USA*

**Abstract.** Developing a general theory of cognition based on formal notions of information remains a long-term goal. One means of making incremental progress toward this goal is to analyze core cognitive capacities to determine whether they can be explained by reference to information. Chunking is one of the most general and least understood phenomena in human cognition. George Miller described chunking as "a process of organizing or grouping the input into familiar units or chunks." The psychological literature describes chunking in many experimental situations but it says nothing about the intrinsic, mathematical properties of chunks. The cognitive science literature discusses algorithms for forming chunks, each of which provides a kind of explanation of why some chunks rather than others are formed, but there are no explanations of what these algorithms, and thus the chunks they find, have in common. We argue that chunks share a common information-theoretic signature. This signature is defined in terms of the basic measure of information content, entropy: Chunks have low conditional entropy internally, and high conditional entropy at the boundaries. We explain this chunk signature and examine several lines of evidence that support this information-theoretic view of chunks. The first is that algorithms built to find chunks based on this signature (or very similar signatures) are quite successful at chunking real-world data. The second is that real chunks, such as words in natural language, appear to be nearly optimally constructed with respect to this signature. Empirical studies also suggest that children, even infants, do actually possess such a chunking ability. All of this evidence supports the view that chunks can be defined by an information-theoretic signature, and that a general chunking ability based on this signature provides a good explanation for this core cognitive ability.

### 1. Introduction

Developing a general theory of cognition based on formal notions of information remains a long-term goal. One means of making incremental progress toward this goal is

to analyze core cognitive capacities to determine whether they can be explained by reference to information. Chunking is one of the most general and least understood phenomena in human cognition. George Miller described chunking as "a process of organizing or grouping the input into familiar units or chunks." Other than being "what short term memory can hold 7 +/- 2 of," chunks appear to be incommensurate in most other respects. Miller himself was perplexed because the information content of chunks is so different. A telephone number, which may be two or three chunks long, is very different from a chessboard, which may also contain just a few chunks but is vastly more complex. Chunks contain other chunks, further obscuring their information content. The psychological literature describes chunking in many experimental situations but it says nothing about the intrinsic, mathematical properties of chunks. The cognitive science literature discusses algorithms for forming chunks, each of which provides a kind of explanation of why some chunks rather than others are formed, but there are no explanations of what these algorithms, and thus the chunks they find, have in common.

We argue that chunks share a common information-theoretic signature. This signature is defined in terms of the basic measure of information content, entropy. Entropy measures the average amount of information required to communicate the outcome of a random variable. For example, the entropy of a toss of a fair six-sided die is much higher than that of a loaded one. In entropic terms, the chunk signature is simple: Chunks have low conditional entropy internally, and high conditional entropy at the boundaries. For example, given the sequence "victo", the conditional entropy of the next letter in the chunk is low (it is probably an 'r'), but given the letters in the chunk "victory", the conditional entropy of the neighboring letters is high. This relationship between predictability and the boundaries of words was noticed as early as 1948 by Claude Shannon.

## **2. Supporting Evidence**

There are several lines of evidence that support this information-theoretic view of chunks. The first is that algorithms built to find chunks based on this signature (or very similar signatures) are quite successful at chunking real-world data. Several such algorithms have been developed independently of one other in the fields of computational linguistics and artificial intelligence, adhering to the chunk signature with varying degrees of fidelity. Perhaps the fullest implementation is that of the Voting Experts algorithm originally developed by Cohen and Adams. Variants of this algorithm, that add bootstrapping (the ability to feed information about chunks already discovered back into the algorithm's decision-making process), represent the highest levels of performance in the literature on a common benchmark of unsupervised chunking ability. Interestingly, this benchmark involves finding words in a corpus of transcribed child-directed speech from the CHILDES project. However, performance of the Voting Experts family of algorithms is not restricted to child language data, as these algorithms also perform well at finding words in diverse languages with different writing systems, finding episodes in sequences of robot actions, finding letters on a printed page by analyzing columns of pixels, and finding teaching episode boundaries in the instruction of an AI student.

While this evidence suggests that algorithms searching for the chunk signature very often recover correct chunks, it does not fully establish the correspondence between the chunk signature and real chunks. The question remains whether real chunks are optimal with respect to this signature. Put more simply, out of all the possible chunks that could be formed based on some data, are the true chunks the "chunkiest?" This question is difficult to evaluate because it requires enumerating an exponential number of possible ways to chunk a given sequence. However, for short sequences, it is possible to fully test this proposition. We developed a chunkiness score that combines the internal entropy and the boundary entropy into a single number. For each 5-word sequence in a corpus of child-directed speech, we generated all possible segmentations and ranked each one according to the chunkiness score. The true segmentation ranked in the 98.7th percentile on average. Preliminarily, it appears that syntax is the primary reason that the true segmentation is not higher in the ranking: When the word-order in the training corpus is scrambled, the true segmentation is in the 99.6th percentile. Still, based on these early results we can say that, in at least one domain, true chunks are nearly optimal with respect to the information-theoretic chunkiness score.

Empirical studies also suggest that children, even infants, do actually possess such a chunking ability. Saffran, Aslin, and Newport famously demonstrated that 8-month-old infants can correctly identify artificial words in a continuous speech stream. Importantly, this speech stream did not contain pauses around sentences or phrases as natural speech often does. This means that infants must be relying on some sort of chunking ability to discover these words in the stream. Saffran et al. proposed a very simple chunking heuristic that was sufficient for their task, but fails at finding words in natural languages and other non-linguistic chunking tasks. In our view, positing such a weak ability is not parsimonious because it would require the children to also have a second, more powerful ability for other chunking tasks, even other linguistic tasks. By contrast, with a single chunking ability based on the signature of chunks, children could perform the task presented by Saffran et al. as well as many others. It is also worth noting that Hauser, Newport, and Aslin later showed that cotton-top tamarins can perform a very similar task, suggesting that the underlying ability may be shared with other non-human primates.

### **3. Conclusion**

All of this evidence supports the view that chunks can be defined by an information-theoretic signature, and that a general chunking ability based on this signature provides a good explanation for this core cognitive ability.

## THE DYNAMISM OF INFORMATION ACCESS FOR A MOBILE AGENT IN A DYNAMIC SETTING AND SOME OF ITS IMPLICATIONS

LARS-ERIK JANLERT

*Umeå University*

*lej@cs.umu.se*

Given the definition of *informational distance* as the time it takes to satisfy a request for the information (Janlert, 2006a), it follows that these distances, the latencies of information satisfactions, will depend on the location of the information-seeking agent as well as the location of the various resources available for satisfying requests for information. That also means that changes in the agent's location as well as changes in the location of information resources in the environment of the agent will dynamically affect the agent's information *availability profile* (Janlert 2006a), the spectrum of informational distances for the complete range of possible information requests. This paper will start to investigate the implication this may have for the possibility of outlining the informational boundaries of the agent, separating agent from world in informational terms, and for the possibilities of strategic relocations of agent and informational resources.

To do this a model of agent–world relationship is outlined and used, more general and considerably more abstract than the examples of actual “natural” agent–world relationships found in this world, starting from a characterization as completely as possible in informational terms: the world is basically a database from which the agent gets information *and* in which the agent sets information.

It turns out that it is possible to define the existential extension of an agent in informational terms in a way that at least starts to make some sense in the real world: the informational boundary. The issue of agent identity may then be approached along the lines of Nozick's closest-continuer theory.

Finally, the importance of proximity as a cue to contextual relevance for situated activity in general is transformed or translated to informational terms to appear as a relevant principle in getting as well as in setting information.

Issues of accuracy and reliability of (purported) information will be bracketed off in this paper, but basically “information” is taken to exclude “misinformation.”

### 1. The world as a database

In this model, we have an *agent* in an environment, a (or the) *world*. The agent is *part* of the environment, but other than that nothing is assumed about its structure and extent or

what drives it. What the agent *does* is two things (which may in the end turn out to be one and the same thing at a certain level of abstraction). Firstly, it requests and gets information from the world. The world is considered to be a (dynamic) repository of information from the agent's point of view: all it ever gets is information from it about it. In our *use* of the model we may of course consider any kind of implementation (model) satisfying the constraints of the agent's interactions.

Secondly, and this is in order to make the model as purely informationally based and symmetric as possible, the agent also *sets* information into the world. Thus, the agent gets as well as sets information.

That is the general model. Such worlds could of course be very different but let us assume for the current exercise that the world of the model by and large matches our own real world at a slightly less abstract level.

Setting or getting information can be viewed as a matter of direction of fit. Getting information can be understood in terms of retrieving, computing, measuring, observing etc., and any combination of such processes, which are partly initiated and performed by the agent (Janlert, 2006b). Setting information means to *make* something the case, to make the world deliver certain information. Getting information is often thought of as a non-intervening process supposed to leave the world untouched, whereas setting information, making something the case, usually is thought of as doing some measure of violence to the world, forcing it to change. But generally in this world you can't get information without setting some information in the process; and you can't set information without getting some information in the process.

Situated existence in this model becomes a kind of information management; we are already living in an informational world, if you will.

This whole approach could in itself perhaps be viewed as an analysis in the style of Carnap (1961); it has certainly been inspired by it.

## **2. Informational boundary of an agent**

Given an agent that moves, it will be possible to make a differentiation between information that is moved "along with" the agent, identifiable as information that is reasonably close and whose distance does not vary much during movement, and information that doesn't. (The size of changes should be understood as relative, in proportion to the whole distance.) Information that moves along with the agent in this sense is considered to be within its (current) informational boundary, other information considered to be on the outside.

For information that does not move along, that is external to the informational boundary, it is also interesting to differentiate between information that is far off, far away at the information horizon of the agent, and whose distance remains fairly constant during the movement of the agent. It will appear as a quite stable background. What remains will then be information that is close to "midrange" *and* changes significantly during movement: proximal external information.

### 3. Proximity principle applied to the informational world

*Things that are close tend to matter; things that matter tend to be(come) close* (Janlert, 2003). For an agent situated in an environment this means roughly: (1) that an object close to the agent has a better chance of getting the agent's attention and figure in the agent's activities; (2) an object that matters to the agent's activities, is more likely to already be or soon become within close range (partly due to the agent's own doings). In the world-as-database model this translates to the following rule of thumb for proximal external information: information that is close to the agent has a better chance to be got by the agent and play a role in the agent's activities; information that matters to the agent's activities, is more likely to be or become close to the agent.

### References

- Janlert, L. E. (2003). Contextual strategies – notes for a theory of context. Technical report UMINF 02.23, Umeå University, ISSN-0348-0542.
- Janlert, L. E. (2006a). Available information — preparatory note for a theory of information space. *tripleC* 4(2). ISSN 1726-670X.
- Janlert, L. E. (2006b). Information at a distance. In *Proceedings of iC&P 2006* (Int. Conf. on Computers & Philosophy), Laval, May 2006.
- Carnap, R. (1961). *Der logische Aufbau der Welt*. Hamburg: Felix Meiner Verlag. First edition appeared in 1928.

## CONTEXTUAL INFORMATION

### *Modeling Different Interpretations of the same Data within a Geometric Framework*

KIRSTY KITTO

*Faculty of Science and Technology  
Queensland University of Technology  
Brisbane, 4001, Australia*

**Abstract.** Semantic Information has provided an elegant set of approaches that allow us to ground information with respect to its Context, Level of Abstraction and Purpose. Interestingly, computer science also has a history of considering context and attempting to incorporate it into fields such as Artificial Intelligence, Ubiquitous Computing, Information Systems design etc. These fields generally treat context as an unknown parameter, which tends to be insufficient when it comes to the modeling of cognition. This paper draws attention to a class of contextuality that arises from "knowing too differently" rather than "too little", and discusses the manner in which this new class is likely to be of increasing importance to the modeling of socio-technical and environmental systems. A new geometric model is discussed which incorporates context at its core. Thus, this paper presents an approach that might be used to ground the truth of statements within a relevant context. Such models make the manner in which context can affect the interpretation of information explicit, and can both consistently explain, and allow us to model, an important class of social phenomena. The model will be discussed with reference to both push polling, and the climate change debate.

### **1. Information in Context**

Semantic Information (Floridi 2011) has provided an elegant set of approaches that allow us to ground information with respect to its Context, Level of Abstraction and Purpose, which has in turn allowed Floridi develop a number of theories about truth, relevance, the logic of being informed etc. (Floridi 2011). However, little work has been presented as to how this theory could correspond to the humans to whom it generally refers, and perhaps most importantly, to their aggregate behavior in e.g. elections, social movements and crises. Semantic Information has the potential to shed some light upon the responses exhibited by individuals to many of the complex information environments that surround them, but realistic models will be required before this can be achieved. While it is relatively easy to determine if the beer is in the fridge (or not), recent public debates on climate change, water management, consumer spending habits in the wake of

the global financial crisis etc. have all served to emphasize the manner in which different sections of a community might ascribe very different values to statements generated from highly similar sets of data. The interpretation that should be attached to information is frequently the subject of vigorous debate, in which context tends to play a fundamental and highly complex role. This situation is recognized somewhat in Floridi's (2011) discussion of semantic truth however, the manner in which such a conception might be worked into the computational modeling of social dynamics is yet to be considered. As scientists attempt to construct increasingly sophisticated climate, water and socio-political models, it has become essential that we consider the manner in which humans respond to complex sets of information and data.

This paper will discuss a sophisticated agent based model (ABM) of human decision making *in context* that is currently in development. This model took inspiration from the work of Brugnach et. al (2008), who contrasted the difference between “knowing too little” a concept already extensively discussed in the computational literature (Akman & Surav 1996, Brézillon 1999), and “knowing too differently”, a concept which is yet to be incorporated into the computational paradigm. To “know too differently” implies a contextual dependency to knowledge, which must be accounted for in models of human behavior.

Taking a situation of water shortage as an example, it is frequently the case that a number of different framings can be provided. This results in the attribution of different interpretations to the situation, each potentially requiring different responses; how should a government react? A farmer will be concerned with “insufficient supply”, while environmentalists might approach the water system thinking that the problem is one of “excessive consumption” (Brugnach et. al 2008). Both contexts have led to claims that are justified, but the two interpretations are incompatible, in that they apparently require different actions from policy makers.

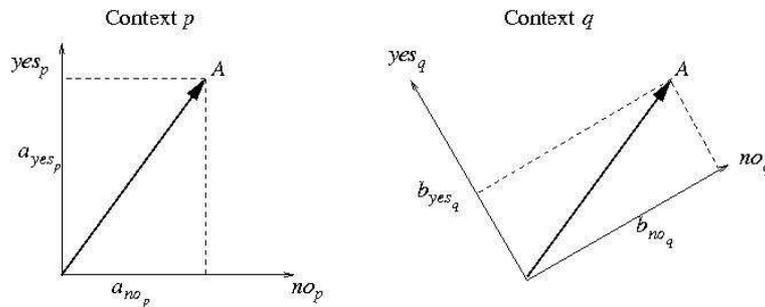


Figure 1. The changing context of a decision. The probability of choosing a particular course of action changes between contexts  $p$  and  $q$ .

While relativistic arguments have a somewhat dubious reputation in pure philosophy, it is becoming increasingly important that we recognize the role context plays in the modeling of human responses to information, and in particular, to the decisions that humans make in utilizing this information. For example, when presented with the same set of information, a different individual might draw a very different set of conclusions as to its consequence, and this can in turn lead to markedly different actions.

The manner in which the new model represents context is geometrical, and can be quickly explained with reference to the simple example illustrated in Figure 1. Here, we have represented the current state,  $A$ , of an agent (we shall call her Alice) with respect to two different contexts  $p$  and  $q$ . In this case, the state of our agent has been chosen to correspond to her projected response to a binary question e.g. will you vote for candidate  $X$  in the coming election?

A connection to probability is generated by assuming that the length of the state  $A$  is equal to 1, which means that the probabilities of Alice responding with a “yes” or “no” are given by the Pythagoras theorem *in a particular context*. Thus,

$$1 = P_{yes} + P_{no} = \begin{cases} |a_{yes_p}|^2 + |a_{no_p}|^2 & \text{in context } p \\ |b_{yes_q}|^2 + |b_{no_q}|^2 & \text{in context } q \end{cases} \quad (1)$$

With reference to Figure 1, it can quickly be seen that the probability of Alice responding with “yes” will be markedly different between the two contexts; while she has a higher probability of responding with “yes” in context  $p$ , she has a higher probability of responding with a “no” to the same question in context  $q$  (this is given by a quick inspection of the lengths of the components making up a right angled triangle with hypotenuse equal to state  $A$ ).

This geometric model of decision making in context bears a remarkable resemblance to the geometrical probability that is utilised in quantum theory (Isham 1995), and indeed, this similarity is further developed in a number of recent contextual models of, for example, decision making (Busemeyer et al. 2011), word recognition and recall (Bruza et al. 2009), concept combination (Aerts & Gabora 2005) and information retrieval (Van Rijsbergen 2004). The general framework of these models will be discussed, and the novel manner in which they incorporate context into the modeling of a state of affairs highlighted. In particular, this paper will highlight the way in which explicitly considering contextual factors in a model allows for a recognition of different points of view and frames *without* lapsing too deeply into relativism. While some notion of truth can be understood to exist in this model, the context in which a set of facts is presented can profoundly influence the interpretation that an agent would attribute to them.

## Acknowledgements

Supported by the Australian Research Council Discovery grant DP1094974.

## References

- Aerts, D. and Gabora, L. (2005). A theory of concepts and their combinations I: the structure of the sets of contexts and properties. *Kybernetes*, 34:151-175.
- Akman, V. and Surav, M. (1996). Steps toward Formalizing Context. *AI Magazine*, 17(3):55-72.
- Brézillon, P. (1999). Context in problem solving: a survey. *Knowledge Engineering Review*, 14:47-80.

- Brugnach, M., Dewulf, A., Pahl-Wostl, C., and Taillieu, T. (2008). Toward a relational concept of uncertainty: about knowing too little, knowing too differently, and accepting not to know. *Ecology and Society*, 13(2):30.
- Bruza, P., Kitto, K., Nelson, D., and McEvoy, C. (2009). Is there something quantum-like about the human mental lexicon? *Journal of Mathematical Psychology*, 53:362-377
- Busemeyer, J. R., Pothos, E., and Franco, R. (2011). A quantum theoretical explanation for probability judgment 'errors'. *Psychological Review*. In press.
- Floridi, L. (2011). *The Philosophy of Information*. Oxford University Press.
- Fox, J. S. (1997). Push Polling: The Art of Political Persuasion. *Florida Law Review*, 49:563.
- Isham, C. J. (1995). *Lectures on Quantum Theory*. Imperial College Press, London.
- Van Rijsbergen, C. (2004). *The Geometry of Information Retrieval*. Cambridge University Press.

## **COGNITION AS MANAGEMENT OF MEANINGFUL INFORMATION. PROPOSAL FOR AN EVOLUTIONARY APPROACH.**

CHRISTOPHE MENANT

### **Extended Abstract**

Humans are cognitive entities. Our behaviors and ongoing interactions with the environment are threaded with creations and usages of meaningful information, be they conscious or unconscious. Animal life is also populated with meaningful information related to the survival of the individual and of the species. The meaningfulness of information managed by artificial agents can also be considered as a reality once we accept that the meanings managed by an artificial agent are derived from what we, the cognitive designers, have built the agent for.

This rapid overview brings to consider that cognition, in terms of management of meaningful information, can be looked at as a reality for animal, humans and robots. But it is pretty clear that the corresponding meanings will be very different in nature and content. Free will and self-consciousness are key drivers in the management of human meanings, but they do not exist for animals or robots. Also, staying alive is a constraint that we share with animals. Robots do not carry that constraint.

Such differences in meaningful information and cognition for animal, humans and robots could bring us to believe that the analysis of cognitions for these three types of agents has to be done separately. But if we agree that humans are the result of the evolution of life and that robots are a product of human activities, we can then look at addressing the possibility for an evolutionary approach at cognition based on meaningful information management. A bottom-up path would begin by meaning management within basic living entities, then climb up the ladder of evolution up to us humans, and continue with artificial agents.

This is what we propose to present here: address an evolutionary approach for cognition, based on meaning management using a simple systemic tool.

We use for that an existing systemic approach on meaning generation where a system submitted to a constraint generates a meaningful information (a meaning) that will initiate an action in order to satisfy the constraint (Menant 2003, 2010 a). The action can be physical, mental or other.

This systemic approach defines a Meaning Generator System (MGS). The simplicity of the MGS makes it available as a building block for meaning management in animals, humans and robots.

Contrary to approaches on meaning generation in psychology or linguistics, the MGS approach is not based on human mind. To avoid circularity, an evolutionary approach has to be careful not to include components of human mind in the starting point

The MGS receives information from its environment and compares it with its constraint. The generated meaning is the connection existing between the received information and the constraint. The generated meaning is to trigger an action aimed at satisfying the constraint. The action will modify the environment, and so the generated meaning. Meaning generation links agents to their environments in a dynamic mode. The MGS approach is triadic, Peircean type.

The systemic approach allows wide usage of the MGS: a system is a set of elements linked by a set of relations. Any system submitted to a constraint and capable of receiving information from its environment can lead to a MGS. Meaning generation can be applied to many cases, assuming we identify clearly enough the systems and the constraints. Animals, humans and robots are then agents containing MGSs. Similar MGSs carrying different constraints will generate different meanings. Cognition is system dependent.

We first apply the MGS approach to animals with “stay alive” and “group life” constraints. Such constraints can bring to model many cases of meaning generation and actions in the organic world. However, it is to be highlighted that even if the functions and characteristics of life are well known, the nature of life is not really understood. Final causes are difficult to integrate in our today science. So analyzing meaning and cognition in living entities will have to take into account our limited understanding about the nature of life. Ongoing research on concepts like autopoiesis could bring a better understanding about the nature of life (Weber and Varela 2002).

We next address meaning generation for humans. The case is the most difficult as the nature of human mind is a mystery for today science and philosophy. The natures of our feelings, free will or self-consciousness are unknown. Human constraints, meanings and cognition are difficult to define. Any usage of the MGS approach for humans will have to take into account the limitations that result from the unknown nature of human mind. We will however present some possible approaches to identify human constraints where the MGS brings some openings in an evolutionary approach (Menant 2010 b & c). But it is clear that the better human mind will be understood, the more we will be in a position to address meaning management and cognition for humans. Ongoing research activities relative to the nature of human mind cover many scientific and philosophical domains (Philpapers, Philosophy of Mind).

The case of meaning management and cognition in artificial agents is rather straightforward with the MGS approach as we, the designers, know the agents and the constraints. In addition, our evolutionary approach brings to position notions like artificial constraints, meaning and autonomy as derived from their animal or human source.

We also highlight that cognition as management of meaningful information by agents goes beyond information and needs to address representations which belong to the central hypothesis of cognitive sciences.

We define the meaningful representation of an item for an agent as being the networks of meanings relative to the item for the agent, with the action scenarios involving the item. Such meaningful representations embed the agents in their environments and are far from the GOFAI type ones (Menant 2010 b). Meanings, representations and cognition exist by and for the agents.

We finish by summarizing the points presented and highlight some possible continuations.

## References

- Menant, C. (2003). Information and Meaning. In: Entropy 2003, 5 (pp193-204). ISSN 1099-4300 © 2003 by MDPI (<http://cogprints.org/3694/>)
- Menant, C. (2010 a). Introduction to a Systemic Theory of Meaning. (short paper) <http://crmenant.free.fr/ResUK/MGS.pdf>
- Weber, A. and Varela, F. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. In: Phenomenology and the Cognitive Sciences 1. (pp 97-125).
- Menant, C. (2010 b). Computation on Information, Meaning and Representations. An Evolutionary Approach. In: Dodig Crnkovic, G. and Burgin, M. (Editors) World Scientific Series in Information Studies - Vol. 2. INFORMATION AND COMPUTATION. Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation. (<http://www.idt.mdh.se/ECAP-2005/INFOCOMPBOOK/CHAPTERS/10-Menant.pdf>.)
- Menant, C. (2010 c). Proposal for a shared evolutionary nature of language and consciousness. <http://cogprints.org/7067/>.
- Philpapers. Philosophy of mind. <http://philpapers.org/browse/philosophy-of-mind>.

## COMPUTATIONAL AND HUMAN MIND MODELS

FRANCISCO HERNÁNDEZ-QUIROZ

UNAM

*Departamento de Matemáticas, Facultad de Ciencias, Ciudad Universitaria, C.P. 04510, D.F., MEXICO*

**Abstract.** Computational models of the human mind have been the subject of a heated debate since Turing's seminal paper of 1950. Some opponents of the so-called Strong AI have postulated alternative mechanisms based on one or another form of hypercomputation. Although specific arguments can be (and have been) raised against the possibility of hypercomputation, a different approach is possible: accept the possibility of human cognitive abilities beyond the reach of Turing Machines (TMs) and then face the problem of postulating appropriate physical mechanisms underlying these hypercomputing abilities. The result can lead to difficulties as hard as those faced by Strong AI in the first place, reducing the allure of the hypercomputing alternatives.

### 1. Introduction

In his celebrated paper of 1950, Turing advanced the then daring proposal of machines able to emulate the human mind. Those machines were the practical realization of the model he introduced before in 1936-7. Turing's formulation is careful to avoid the categorical statement that the human mind can be emulated by a Turing Machine due to the fact that *it is* a Turing Machine. However, successive computer scientists have reprised Turing's proposal without his caveats. An extreme and idealized version of this point of view is known as Strong Artificial Intelligence (Searle, 1984).

### 2. An Objection to Artificial Intelligence

The thesis that the human mind can be modelled by Turing Machines has been attacked by many people. A common line of attack goes like this:

- Strong AI claims the human mind can be modelled by Turing Machines.
- Turing Machines suffer internal limitations that surface in theorems due to Turing himself, Rice and even Gödel.
- But human cognitive abilities go beyond these limitations.
- Ergo, the human mind cannot be modelled by Turing Machines.

This argument has been rejected by many authors (Feferman, 1996; Chalmers, 1995). But this paper will take a different approach: what happens if we accept that the human mind cannot be modelled by a Turing Machine? What type of mechanism is needed instead? What problems arise when such a model is adopted?

### **3. “Mechanisms” more powerful than computers**

There are many candidates for this role. On the one hand, physical systems with properties (supposedly) beyond the restrictions of Turing Machines (Penrose, 1994). On the other hand, mathematical models circumventing those same restrictions: Oracle Turing Machines (Turing, 1939), Analog Neural Networks (Siegelmann, 1999), Dynamical Systems (Bournez and Cosnard, 1995), etc.

In fact, there is a common core in all these models: (a) they pretend to implement some notion of what can be considered intuitively a computational mechanism; (b) simultaneously, they include elements capable of introducing entities not Turing computable. They can be gathered under the label of “hypercomputation.”

Many of those who oppose the Strong AI, claim that human cognitive abilities which are not explicable by TMs are in fact based on one or another hypercomputing mechanism.

### **4. Towards a new scientific research program?**

But these mechanisms are also prone to run into trouble. Sieg (2008) has argued convincingly that Turing Machines' limitations are a consequence of the acceptance of two principles: locality and boundedness. The first principle means that a computer can only change immediately recognizable configurations in finite time. The second one means that a computer can only recognize immediately only a bounded number of configurations (and therefore there exists an upper bound to the amount of information it can handle in finite time).

By rejecting TMs as an upper bound to computability, we reject these principles. No need to worry though, theoretically speaking, if we are only interested in abstract mathematical models. But if the aim is to model or to explain the human mind, and some of its capabilities are attributed to hypercomputing features, then we are asserting implicitly that the human mind (or its physical substratum, if you will) goes beyond the principles of locality and boundedness. One variety of hypercomputation even asserts the possibility of harnessing and manipulating non-computable irrational numbers (Siegelmann, 1999). And if we want to remain on scientific grounds, we will be pressed to point out to the physical counterparts of this theoretical entities and postulate hypercomputation in Nature.

Of course, none of this is impossible, at least in principle. However, our quest for a model of the human mind has lead us to pose very basic questions about physical reality that bring with them huge theoretical and practical challenges that look at least as difficult as the problems faced by the computational models of the human mind. The moral might be that a theoretical alternative is not necessarily a plausible explanation for a natural phenomenon.

## References

- Bournez, O. y Cosnard, M. (1995). *On the computational power and super-Turing capabilities of dynamical systems*, Technical report no. 95-30, Laboratoire de l'Informatique du Parallelism, Ecole Normale Supérieure de Lyon.
- Chalmers, D.J. (1995). Minds, Machines, And Mathematics - A Review of Shadows of the Mind by Roger Penrose", *Psyche* 2(9).
- Feferman, S. (1996). Penrose's Gödelian argument, *Psyche* 2, 21-32.
- Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*, Oxford University Press.
- Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*, Oxford University Press.
- Searle, J. (1984). *Minds, Brains and Science*, Cambridge: Harvard University Press.
- Sieg, W. (2008). Church Without Dogma — axioms for computability. In B. Lowe, A. Sorbi, B. Cooper (eds.) *New Computational Paradigms* (pp. 139-152), Springer Verlag.
- Siegelmann, H.T. (1999) Neural Networks and Analog Computation: Beyond the Turing Limit, Birkhäuser, Progress in Theoretical Computer Science.
- Turing, A.M. (1936-7). On Computable Numbers, with an Application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society*, Series 2, 42, pp. 230–265.
- Turing, A.M. (1939). Systems of Logic Based on Ordinals, *Proceedings of the London Mathematical Society*, Series 2, 45, 161–228.
- Turing, A.M. (1950). Computing Machinery and Intelligence, *Mind*, 59, 433–460.

## SEMANTICS OF INFORMATION

### *Meaning and Truth as Relationships between Information Carriers*

MARCIN J. SCHROEDER

*Akita International University*

*Akita, Japan*

**Abstract.** The meaning of information has been openly dismissed from the interest of information theory already by Shannon, but the fiasco of the early attempt to develop semantic theory of information by Bar-Hillel and Carnap was even more discouraging. They developed their theory of semantic information using as a starting point already existing logical structure of the language, not recognizing the fact that language is a very special information system and the logic of information should be built before its semantic theory. Philosophical concept of meaning for centuries has been associated with the medieval scholastic concept of intentionality, pointing by a symbol at intended object, identified by Brentano and his followers as the primary characteristic of mental acts. Neither of the attempts to eliminate psychologism of intentionality removed the primary source of philosophical problems which has been always in the fact that semantics requires crossing the border between different ontological entities. This difficulty could not be resolved within philosophy of language, as at this level the difference between linguistic items and entities to which they refer cannot be ignored. The relationship between a symbol and its meaning does not require separation of ontological status, when the meaning is understood as a relationship between information in two different information carriers, that of a symbol and that of denotation. In the present paper, both, symbol and object are described in terms of information integration. Every entity is being characterized through the integrated part of information constituting its identity, and not integrated interpreted as its state. The correspondence of identities, i.e. integrated parts of information is here identified as the meaning, the correspondence between states, i.e. nonintegrated parts of information is identified as the truth.

### **1. Sources of Problems in Semantics of Information**

Difficulties in the development of semantics of information are in part inherited from linguistic semantics, but some of them have their sources in the circumstances in which information theory has been born. The meaning of meaning has been always an elusive subject. Ogden and Richards (1923/1989) in their widely read study of this concept considered its sixteen basic meanings.

Philosophical concept of meaning for centuries has been associated with the medieval scholastic concept of intentionality, pointing by a symbol at intended object.

Brentano identified intention or “aboutness” with the fundamental characteristic of mental capacity.

The logical approach initiated by Frege and developed by Church was an attempt to eliminate psychological aspects of the meaning by making a distinction between denotation and sense, and focusing on the rules reducing sense of compound expressions to those simple. However, the shift of attention to mutual relationship between expressions of a language at different level of complexity does not help to understand the relationship between simple signs and their denotations, to which the process of reduction is leading. Under influence of logical positivism, Carnap attempted to resolve this issue in the context of scientific methodology by involving the idea of empirical sense reducing criteria of the relationship to empirical procedures.

The approach initiated by Peirce, whose original writings preceded most of the contemporary work on the concept of meaning, was also intended as a way to eliminate necessity to involve human subject in semiosis. In his approach sign and object are accompanied by interpretant of the type of a sign. Being a sign, interpretant may enter into another triadic relation with its own object and interpretant. Its role is to build a connection between sign and object which does not require involvement of human being. This approach leaves the question of the traditional relationship between the sign and its meaning open-ended, but it hardly gives its explanation, especially when the sign has different ontological status from that of an object. As in the logical approach, we have here an extension of the study towards a complex structure of signs or names, but the basic relationship between the object and the sign is left in the shadow.

No wonder that the issue of the meaning of information has been dismissed from the subject of information theory so easily. Shannon’s disclaimer “These semantic aspects of communication are irrelevant to the engineering problem” (Shannon & Weaver, 1949/1989) has been followed by majority of information theorists, such as Cherry (1951/1952): “It is important to emphasize, at the start, that we are not concerned with the meaning or the truth of messages; semantics lies outside the scope of mathematical information theory.” After all, the measure of information was defined for one letter or character of a message which does not carry any meaning. The measure for entire message was simply the sum of measures for characters.

Fiasco of the early attempts to develop semantic theory of information, such as the most advanced attempt by Bar-Hillel and Carnap (1952), sealed the fate of the study of semantics of information. Bar-Hillel and Carnap developed their theory of semantic information using as a starting point already existing logical structure of the language. They did not take into account that language is a very special information system and more general logic of information should be built before its semantic theory.

## **2. Semantics as Relationship between Information Carriers**

Bar-Hillel and Carnap (1952) have built their measure of semantic information in such a way that it can be reduced to Shannon’s entropy in a special case. However, here there is a fundamental problem whether the measure of information transmitted in the process of communication applies to information carried by some carrier (symbol or object). The present author (Schroeder, 2004) believes that the answer is negative, and the measure of semantic information should be based on the alternative measure, taking into

consideration the amount of information carried by symbols, which should be estimated based on the relationship between the information in the symbol and information in the designate.

However, the primary source of philosophical problems of semantics has been always in the requirement of crossing the border between different ontological entities. This difficulty could not be resolved within philosophy of language, as at this level the difference between linguistic items and entities to which they refer cannot be ignored.

The relationship between a symbol and its meaning does not require separation of ontological status, when the meaning is understood as a relationship between information in two different information carriers, that of a symbol and that of denotation. In the present paper, both, symbol and object are described in terms of information integration (Schroeder, 2009).

The concept of information integration is implemented with the use of a theoretical instrument called a generalized Venn gate which transforms selective manifestation of information into structural one (Schroeder, 2005, 2007) The transition may change the level of integration of information depending on the structural characteristics of the logic of the gate. The gates whose logic is completely irreducible into the components (such as in the case of quantum logic) produce highest level of integration. The gates with Boolean (i.e. traditional) logic reducible to the product of simple (yes-no) components leave information completely disintegrated. There are of course multiple levels of integration in between.

Information is here understood in a very broad way as an identification of a variety, i.e. that which makes one out of a variety (Schroeder, 2005). Thus, not only language is a carrier of information, but also every object of our experience. Cognitive processes involve transformations of selective manifestation of information coming with sensory stimulation into the structural manifestation of information, which in its integrated form constitute conscious experience.

Every entity is being characterized through the integrated part of information constituting its identity, and not integrated interpreted as its state. The correspondence of identities, i.e. integrated parts of information is here identified as the meaning, correspondence between states, i.e. non-integrated parts of information is identified as the truth.

## References

- Bar-Hillel, Y. & Carnap R. (1952/1964). An Outline of a Theory of Semantic Information. Technical Report No. 247, Research Laboratory of Electronics, MIT; reprinted in Bar-Hillel, Y. (1964) *Language and Information: Selected essays on their theory and application*. Reading, MA: Addison-Wesley, pp. 221-274.
- Cherry, E. C. (1951/1952). A history of the theory of information. *Proceedings of the Institute of Electrical Engineers*, 98 (III), 383-393; reprinted with minor changes as: The communication of information. *American Scientist*, 40, 640-664.
- Ogden, C. K., Richards, I. A. (1923/1989). *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism*. San Diego: A Harvest Book, Harcourt Brace Jovanovich.
- Schroeder, M. J. (2004). An Alternative to Entropy in the Measurement of Information. *Entropy*, 6, 388-412.

- Schroeder, M. J. (2005). Philosophical Foundations for the Concept of Information: Selective and Structural Information. In *Proceedings of the Third International Conference on the Foundations of Information Science, Paris*. <http://www.mdpi.org/fis2005>.
- Schroeder, M. J. (2007). Logico-algebraic structures for information integration in the brain. *Proceedings of RIMS 2007 Symposium on Algebra, Languages, and Computation*, Kyoto: Kyoto University, pp. 61-72.
- Schroeder, M. J. (2009). Quantum Coherence without Quantum Mechanics in Modelling the Unity of Consciousness. In P. Bruza, et al. (Eds.) *QI 2009, LNAI 5494*, Springer, pp. 97-112.
- Shannon, C. E., Weaver, W. (1949/1998). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

## PRE-COGNITIVE SEMANTIC INFORMATION<sup>6</sup>

ORLIN VAKARELOV  
*Department of Philosophy*  
*University of Arizona*  
*Tucson, Arizona, USA*  
*Email: okv@u.arizona.edu*

**Abstract.** This talk addresses one of the fundamental problems of the philosophy of information: How does semantic information emerge within the underlying dynamics of the world? --- dynamical semantic information problem. It suggests that the canonical approach to semantic information that defines data before meaning and meaning before use is inadequate for pre-cognitive information media. Instead, we should follow a pragmatic approach to information where one defines the notion of information system as a special kind of purposeful system emerging within the underlying dynamics of the world, and define semantic information as the currency of the system. In this way, systems operating with semantic information can be viewed as patterns in the dynamics – semantic information is a dynamical system phenomenon of highly organized systems. In the simplest information systems the syntax, semantics and pragmatics of the information medium are co-defined. It proposes a new more general theory of information semantics that focuses on the interface role of the information states in the information system – the interface theory of meaning.

### 1. Introduction

I address the following problem: How does semantic information emerge within the underlying dynamics of the world? Let us call this the dynamical semantic information (DSI) problem. This is related to another kind of problem: Can we provide a foundation of cognitive science with the notion of (semantic) information? I claim that it is possible to offer a theory of pre-cognitive semantic information that does not presuppose a notion of cognition or mind. With such a theory, the notion of semantic information can be used in foundational discussions of cognition without circularity. However, I do not plan to address the second problem here.

My strategy for addressing DSI is this: Start with a notion of *information system* that is a special kind of autonomous dynamical system interacting with an environment. Describe semantic information as a “currency” of the information system. That is, treat information for the system not as a primitive but as a derived notion, similar to the way currency is a derived notion of an economic system. Take a *decomposition approach* to

---

<sup>6</sup> This talk is based on (Vakarelov, 2010).

analyzing the components of semantic information – that is, regard notions such as data, meaning and source, as depicting aspects of informational processes within the information system. Provide a theory of meaning, the *interface theory of meaning*, for the informational states of an information medium within the information system.

## 2. Canonical Views of Semantic Information

Most theories of semantic information make the following assumptions: (1) semantic information = data + meaning (+ truthfulness); (2) the data is conceptually primary; (3) meaning is secondary and depends on data, (4) pragmatics is third-ary and depends on meaning. In this view, the ‘+’ in the definition of information can be regarded as an amendment operation, where syntax is amended by semantics to obtain a theory of semantic information, and semantics is amended with an account of use of information, to obtain a theory of pragmatic information. Thus, an approach to semantic information proceeding as such I call an *amendment approach*.

Taking an amendment approach to semantic (and pragmatic) information has no effect on the formal theories of information; however it affects meta-theoretic judgments about theories of information. In particular, it affects what theories of information are regarded as more general.

I argue (defeasibly) that taking the notion of *data* as conceptually primary (and independent from semantics and pragmatics) leads to an indispensable role of a mind for the specification of semantics. This makes naturalizing semantic information difficult. This is because the cases where the data system can be defined precisely without semantics or pragmatics are cases where semantics requires an external interpreter. The meta-theoretical judgments about such cases mistakenly conclude that the cases are the most general, and therefore they offer the most inclusive theory of semantic information.

## 3. The Pragmatic Approach to Semantic Information

I propose an alternative: I argue for a decomposition approach to information; that is, I argue that in the most general case of semantic information, *data*, *semantics*, and *pragmatics* are codetermined as aspects of an information process. The most general kind of information is pragmatic information; that is, in the most general case, semantic information requires a system that utilizes information in its interaction with an environment. Such a system I call, following (Nauta, 1997), an *information system*.

The strategy of pragmatic analysis of information is the following: The most basic notion is *information system*. An information system *S* is a physical system that is in an active interaction with an external environment and that satisfies a set of conditions that do not presuppose the notion of information. The conditions must guarantee the existence in *S* of a sub-system, *M*, that can be interpreted as an information medium. Moreover, the functional role of *M* in *S* in relation to the interaction with the environment must be sufficient to define the semantic content of the states of *M*.

According to this strategy, *S* is an information system not because it operates with meaningful information, but conversely, it operates with information because it is an information system. The most important idea is that what counts as data, and what gives the data semantic content, is determined by the role they play in the information system.

#### 4. Information Systems

An information system  $S$  is a system that satisfies the following five conditions:

1.  $S$  is an *open system*, i.e. it is a system that is distinct from its environment, but it is in constant interaction with the environment.
2.  $S$  is a *partially isolated* open system, i.e. some of the interactions between  $S$  and the environment are structured through well-defined limited channels of influence.
3.  $S$  is a *purposeful system*. That is, there is at least one proper set of goal states,  $G$ , that the system “attempts” to be in (or near) by affecting its environment.
4.  $S$  contains a sub-system  $M$  that can correlate with an external system  $O$ , and  $M$  can control the behavior of  $S$ .
5.  $S$  contains a second distinct sub-system  $P$  that filters the states of  $M$  and their effect on behavior in relating to its purpose. In other words,  $P$  steers the system towards  $G$  by modulating the control effect of  $M$ .

I argue that all the conditions for an information system can be depicted (in *principle*) as conditions of dynamical systems. Thus, no mentalistic or cognitive notions are needed to define an information system. I also argue that the conditions are sufficient to justify regarding  $M$  as an information medium with states that can be interpreted as data/information states, and as having meaning for the system. The data/information states of  $M$ , however depend on the global dynamics. In particular, they depend on the way  $P$  modulates the control function of  $M$  and on the states of  $O$  (which can be regarded as an information source). However, the states of  $O$  and  $P$  also depend on the global dynamics. Thus, in the most general information systems all relevant components of the information system are codetermined (except the goal  $G$ ).

#### 5. Interface Theory of Meaning

In an information system content is determined neither by the external relation between  $M$  and  $O$ , nor by the internal role of the states of  $M$  in  $S$ , but by the *interface roles* the states of  $M$  play in the dynamics of the system. This is the interface theory of meaning for information states in an information system. More traditional theories of semantics, such as correspondence semantics or conceptual role semantics, can be obtained from the interface role semantics as aspects of the interface relation. Thus, the interface theory of meaning properly generalizes other theories of meaning, which only work if further conditions on the information system are demanded.

#### References

- Nauta, D. (1970). *The Meaning of Information*. The Hague: Mouton.
- Vakarelov, O. (2010). Pre-cognitive semantic information. *Knowledge, Technology & Policy*, 23(1):193-226.

# **Track III: Autonomous Robots and Artificial Cognitive systems**

## WHO WILL HAVE IRRESPONSIBLE, UNTRUSTWORTHY, IMMORAL INTELLIGENT ROBOT?

*Why Artificially Intelligent Adaptive Autonomous Agents need to be Artificially Moral?*

MARGARYTA GEORGIEVA ANOKHINA

*School of Innovation, Design and Engineering, Mälardalen University, Sweden [maa05002@student.mdh.se](mailto:maa05002@student.mdh.se)*

and

and Gordana Dodig Crnkovic

*School of Innovation, Design and Engineering, Mälardalen University, Sweden [gordana.dodig-crnkovic@mdh.se](mailto:gordana.dodig-crnkovic@mdh.se)*

**Abstract.** We argue that there is natural place for artificial moral agency parallel to artificial intelligence.

### 1. Extended Abstract

Historically, moral agency was conceptualized in purely anthropocentric terms. Consequently, only humans qualify as moral agents according to the traditional criteria and no other agents than humans were considered capable of moral agency. We discuss such conventional criteria as mental states, intentionality, autonomy, free will, responsibility, rationality and moral reasoning and compare human agents with artificial agents (intelligent adaptive learning robots and software agents, present and envisaged in coming decades).

We attempt to understand what has shaped traditional criteria in the past and how technological change initiates re-shaping the world around us, including what we could (and should) be considered as moral agents.

We suggest that conventional approach to moral agency is unable to provide exhaustive criteria to deal with moral situations of contemporary world involving techno-social systems with autonomous intelligent agents, both humans and artifacts. We also discuss how morality can be approached in new ways in case of artificial agents. The argument is provided that human-centric approach to intelligent autonomous machines is inappropriate as a means of control of behavior in self-learning artificial agents and a

new proposal is made about how to treat notion of moral responsibilities in techno-social systems when intelligent artifacts acting autonomously are involved.

In the past mechanical age of engineering, technological systems were designed to perform specific and limited functions and they were kept closed with no access to the outside world (like a robot making car parts, for example). Nowadays systems with artificial intelligence are more complex and sophisticated and they are starting to be implemented in everyday environments like people's homes in helping elderly and sick people and as companions (the developing field of social robotics).

This rapid technological change re-shapes and expands ways of thinking about agency and morality that we used to have. Machine "talks", "selects", "runs" "reasons", "senses", "plays chess", etc. not in a human way, but we use these words to express functionality of a machine in familiar terms. Why can't machine "choose", "decide", "think" or "be responsible"?

In the similar way as machines are artifactually intelligent, they can be and indeed must be made artifactually moral if we are to rely on them even when they are not under direct control, when they act autonomously. The term "artificial intelligence" reveals the same problem one had to accept that machine can behave intelligently even though it is intelligence of an artifact, and not a human intelligence.

Similarly, machine can be made functionally, artificially moral. It may take some effort to find out how to secure morally acceptable behavior in intelligent learning machines, and some researchers suggest it may take as much effort as it took for the development of artificial intelligence. But it would be irresponsible to let them go among people without having morally acceptable behavior according to human standards.

Floridi and Sanders (2004) consider interactivity, autonomy and adaptability at a given level of abstraction as important new criteria for moral agency. Morality in this approach is thought of as "a threshold defined on the observables in the interface". These criteria are related to criteria of operational environment, suggested by Berthier (2006) and domain, suggested by Foner (1993). This requirement relates to differences between domains of interest for moral considerations for human agents and for artificial ones. As humans act and behave in specific environment, artificial agents do as well, but conditions are different, and thus probably not all criteria that are suitable for human domain are applicable to operational environment of artificial agents. Both artificial agents and humans need interaction and ability to adapt to environment in order to act morally, according to the rules that define moral actions. Coeckelbergh (2009) suggests using the term *virtual morality*, as robots can exhibit behaviour akin to behaviour of humans in analogous situations.

The aim of the emerging research field of machine ethics (machine morality, artificial morality, or computational ethics) such as developed in Anderson and Anderson (2007); Allen, Wallach, Smit (2006) and Moor (2006) is moral decision-making implemented in computers and robots.

We discuss parallels between artificial agent's possible artifactual moral agency, see Dodig-Crnkovic and Persson (2008), similarity and differences compared to human agents. We argue that there is natural place for artificial moral agency parallel to artificial intelligence.

## References

- Floridi L. and Sanders J. W. (2004) On the Morality of Artificial Agents. *Minds and Machines* 14 (3):349-379.
- Berthier D. (2006) Artificial Agents and their Ontological Status, *iC@P 2006: International Conference on Computers and Philosophy*, p.2-5.
- Foner L. (1993) What's An Agent, Anyway? A Sociological Case study, available from the Agents Group, MIT Media Lab.  
<http://www.nada.kth.se/kurser//kth/2D1381/JuliaHeavy.pdf>, p.35.
- Coeckelbergh M. (2009) Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents, *AI & Soc* 24:188-189.
- Anderson M. and Anderson S. L. (2007) Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine* Volume 28 Number 4.
- Allen C., Wallach W., Smit I. (2006) Why Machine Ethics?, *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 12-17, July/Aug. 2006, doi:10.1109/MIS.2006.83.
- Moor J. H. (2006) The Nature, Importance, and Difficulty of Machine Ethics, *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 18-21, July/Aug. 2006.
- Dodig-Crnkovic G. and Persson D. (2008) Sharing Moral Responsibility with Robots: A Pragmatic Approach. Tenth Scandinavian Conference on Artificial Intelligence SCAI 2008. Volume 173, *Frontiers in Artificial Intelligence and Applications*. Eds. A. Holst, P. Kreuger and P. Funk.

## THE ETHICS OF ROBOTIC DECEPTION

*RONALD C. ARKIN*

*Mobile Robot Laboratory, Georgia Institute of Technology  
85 5th ST NW, Atlanta, GA 30332 U.S.A.*

The time of robotic deception is rapidly approaching. While there are some individuals trumpeting about the inherent ethical dangers of the approaching robotics revolution (e.g., Joy, 2000; Sharkey, 2008), little concern, until very recently, has been expressed about the potential for robots to deceive human beings. Our working definition of deception (for which there are many) that frames the rest of this discussion is “deception simply is a false communication that tends to benefit the communicator” (Bond and Robinson, 1988). Research is slowly progressing in this space, with some of the first work developed by Floreano et al (2007) focusing on the evolutionary edge that deceit can provide among an otherwise homogeneous group of robotic agents. This work did not focus on human-robot deceit, however. As an outgrowth of our research in robot-human trust (Wagner and Arkin, 2008), where robots were concerned as to whether or not to trust a human partner rather than the other way around, we considered the dual of trust: deception. As any good conman knows, trust is a precursor for deception, so the transition to this domain seemed natural. We were able to apply the same models of interdependence theory (Kelley and Thibaut, 1978) and game theory, to create a framework whereby a robot could make decisions regarding both when to deceive (Wagner and Arkin, 2009) and how to deceive (Wagner and Arkin, 2011). This involves the use of partner modeling or a simplistic view (currently) of theory of mind to enable the robot to (1) assess a situation; (2) recognize whether conflict and dependence exist in that situation between deceiver and mark, which is an indicator of the value of deception; (3) probe the partner (mark) to develop an understanding of their potential actions and perceptions; and (4) then choose an action which induces an incorrect outcome assessment in the partner.

While the results we published (Wagner and Arkin, 2011) we believe were modestly stated, e.g., “they do not represent the final word on robots and deception”, “the results are a preliminary indication that the techniques and algorithms described in this paper can be fruitfully used to produce deceptive behavior in a robot”, “much more psychologically valid evidence will be required to strongly confirm this hypothesis”, etc. The response to this research has been quite the contrary, ranging from accolades (being listed as one of the top 50 inventions of 2010 by Time Magazine (Suddath, 2010)) to damnation (“In a stunning display of hubris, the men ... detailed their foolhardy experiment to teach two robots how to play hide-and-seek” (Tiku, 2010), and “Researchers at the Georgia Institute of Technology may have made a terrible, terrible mistake: They’ve taught robots how to deceive” (Geere, 2010)).

It seems we have touched a nerve. How can it be both ways? It may be where deception is used that forms the hot button for this debate. For military applications, it seems clear that deception is widely accepted (which indeed was the intended use of our

research as our sponsor is the Office of Naval Research). Sun Tzu is quoted as saying that “All warfare is based on deception”, and Machiavelli in *The Discourses* states that “Although deceit is detestable in all other things, yet in the conduct of war it is laudable and honorable”. Indeed there is an entire U.S. Army (1988) Field Manual on the subject.

In our original paper (Wagner and Arkin, 2011), we included a brief section on the ethical implications of this research, and called for a discussion as to whether roboticists should indeed engage in this endeavor. In some ways, outside the military domain, the dangers are potentially real. And of course, how does one ensure that it is only used in that context? Is there an inherent deontological right, whereby humans should not be lied to or deceived by robots? Kantian theory clearly indicates that lying is fundamentally wrong, as is taught in most introductory ethics classes. But from a utilitarian perspective there may be times where deception has societal value, even apart from the military (or football), perhaps in calming down a panicking individual in a search and rescue operation or in the management of patients with dementia, with the goal of enhancing that individual’s survival. In this case, even from a deontological perspective, the intention is good, let alone from a utilitarian consequentialist measure. But does that warrant allowing a robot to possess such a capacity?

The point of this paper is not to argue that robotic deception is ethically justifiable or not, but rather to help generate discussion on the subject, and consider its ramifications. As of now there are absolutely no guidelines for researchers in this space, and it indeed may be the case that some should be created or imposed, either from within the robotics community or from external forces. But the time is coming, if left unchecked, you may not be able to believe or trust your own intelligent devices. Is that what we want?

## Acknowledgements

This research was supported by the Office of Naval Research under MURI Grant # N00014-08-1-0696. The author would also like to acknowledge Dr. Alan Wagner for his contribution to this project.

## References

- Bond, C. F., & Robinson, M., (1988). “The evolution of deception”, *Journal of Nonverbal Behavior*, 12(4), 295- 307.
- Floreano, D., Mitri, S., Magnenat, S., & Keller, L., (2007). “Evolutionary Conditions for the Emergence of Communication in Robots”. *Current Biology*, 17(6), 514-519.
- Geere, D., (2010). *Wired Science*,  
<http://www.wired.com/wiredscience/2010/09/robots-taught-how-to-deceive/>
- Joy, B. (2000). “Why the Future doesn’t need us”. *Wired*, April 2000.
- Kelley, H. H., & Thibaut, J. W., (1978). *Interpersonal Relations: A Theory of Interdependence*. New York, NY: John Wiley & Sons.
- Sharkey, N. (2008). “The Ethical Frontiers of Robotics”, *Science*, (322): 1800-1801.
- Suddath, C., (2010). “The Deceitful Robot”, *Time Magazine*, Nov. 11, 2010,  
[http://www.time.com/time/specials/packages/article/0,28804,2029497\\_2030615,00.html](http://www.time.com/time/specials/packages/article/0,28804,2029497_2030615,00.html)
- Tiku, N., (2010). *New York Magazine*, 9/13/2010,  
[http://nymag.com/daily/intel/2010/09/someone\\_taught\\_robots\\_how\\_to\\_1.html](http://nymag.com/daily/intel/2010/09/someone_taught_robots_how_to_1.html)

- U.S. Army (1988). Field Manual 90-2, Battlefield Deception, <http://www.enlisted.info/field-manuals/fm-90-2-battlefield-deception.shtml>
- Wagner, A. and Arkin, R.C., (2008). "Analyzing Social Situations for Human-Robot Interaction", *Interaction Studies*, Vol. 9, No. 2, pp. 277-300.
- Wagner, A. and Arkin, R.C., (2009). "Robot Deception: Recognizing when a Robot Should Deceive", *Proc. IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09)*, Daejeon, KR.
- Wagner, A.R., and Arkin, R.C., (2011). "Acting Deceptively: Providing Robots with the Capacity for Deception", *International Journal of Social Robotics*, Vol. 3, No. 1, pp. 5-26.

**PROLEGOMENON TO ANY FUTURE THEORY OF MACHINE  
AUTONOMY**

PAUL BELLO  
*Office of Naval Research*  
875 N. Randolph St., Arlington VA 22203

AND

Selmer bringsjord  
*Rensselaer Polytechnic Insititute, Dept. of Cognitive Science, Dept. of  
Computer Science, Lally School of Management*  
110 8<sup>th</sup> St, Troy NY 12180

AND

marcello guarini  
*University of Windsor, Dept. of Philosophy*  
401 Sunset Ave., Windsor, Ontario N9B 3P4

**Abstract.** As the development of autonomous systems lead to smarter and more capable machines, we must concern ourselves with the possibility that they will one day be equipped with weapons and the authorization to use them. However, it isn't inconceivable that such systems will be prone to error, leaving us with the issue of who might be to blame if force is misapplied. In this presentation, we discuss responsibility as it pertains to autonomous systems. More specifically, we attempt to give a formal analysis of the conditions under which an autonomous system might consider itself to be a "freely acting agent." Note that we do not attempt to attack the metaphysical problem of free will; we only aim to provide the system with an appropriate commonsense theory of what it means to be free, given a set of circumstances within which the agent acts. Such a commonsense theory will (eventually) contain a set of beliefs corresponding to how external obligations, potential coercion, lack of perfect information, and brute facts constrain or expand the set of actions available to the agent at a given time in branching-time semantics. The semantics represents the agents' beliefs about the past as fixed and the future as a set of possible histories that are contingent on its actions. Future extensions of our formal framework will be discussed relative to the development of a "Moral Turing Test" for autonomous systems.

“You have been terminated.” In grand Hollywood style, this is how much of the public-at-large has been introduced to the notion of autonomous robots on the battlefield. When these words were famously uttered by the now-Governor of California, combat robots were only a dream, and the dystopian future painted in the *Terminator* movies seemed no more imminent than a new ice age. Times have rather changed. Combat robots roam through craggy caves in Afghanistan

searching for terrorists, and unmanned air vehicles strike suspected enemy hideouts in Pakistan without a human operator being anywhere close by. Thankfully, we still live in a pre-Terminator age. The United States Department of Defense maintains strict policies that require humans be in the decision-making loop whenever robots are employed on the battlefield. While this sets many a mind at ease, neither of us are totally convinced that such strictures will indefinitely remain, especially as robots and associated technology becomes more reliable, more intelligent, and --- in the end, the most important factor --- cheaper. Similar scenarios have been discussed at length by (Joy, 2000) and other futurists (Bostrom, 2003). In reply to these concerns, we (Bringsjord, Arkoudas & Bello 2006) and others (Arkin, 2009) have looked to curb robotic behavior through the mechanization of norms, conventions, and other ethical structures, such that future robots might be bound by regulations. Unfortunately, complex situations are the norm on the battlefield, and facing novel moral dilemmas in combat is the rule rather than the exception. Just as our warfighters must improvise under these adverse circumstances, we expect future robots to take actions roughly consistent with pre-established norms, but rounded out with a measure of commonsense moral judgment, for if they do not, they are doomed to be both brittle and ineffectual soldiers.

This being said, we'd like to address an issue at IACAP 2011 that hasn't received much attention in the literature: the issue of whether or not future intelligent robots could be blamed for their actions, provided something goes wrong during the course of their operation. Our plan will be to provide what we feel to be a reasonable set of conditions that when jointly obtaining would allow us to classify a robot as a moral agent, and as such subject to blame in the case of intentional misdoings or derelictions of duty. The key question under consideration in our investigation is: "what does it mean for x to have the property of being autonomous?" We hope to clarify a set of potential confusions about the proper definition of autonomy in the context of robotic warfighters.

Moral philosophers, depending on their particular stance on the nature of morality, typically define autonomy as the ability to respect some particular moral code or another, even if doing so runs contrary to self-interest. In a deep sense, these ideas turn on the notion of an autonomous agent having at least the illusion of free will, or the ability to choose contrary to a pre-established set of normative principles. Among roboticists and other practitioners of artificial intelligence, autonomy has generally been taken to mean the ability to make decisions and take actions without coercion or assistance from a secondary agent. While this seems to be plausible enough, a few mental exercises might convince you that this is much too general, perhaps to the point of not being useful in its intended context.

Consider the case of the lowly thermostat that has functionality allowing it to turn on and off in order to maintain a pre-set ambient temperature in a home. It certainly "makes decisions" about when to turn on, and takes action (e.g. turns on) under an appropriate set of conditions and without consulting an external agent at decision-time. Should this device be granted autonomy? We think not, and we assume that our roboticist colleagues agree with us. Even though the thermostat makes decisions (in some sense) as to when to turn on, it's not at all clear that it could choose otherwise. In fact it cannot, barring device malfunction. Worse than this, there isn't an "it" making decisions at all. It's just a thermostat. If we map

onto the robotic case, it's equally unclear that there is an "it" making decisions, or one making free choices that direct its own affairs.

Real-world battlefield situations don't bifurcate so cleanly when it comes to making moral and non-moral decisions. Simple navigation decisions, such as whether or not to step into a house of worship, seem to be prima facie non-moral in nature, but as we well know, they indeed have moral consequences. These complications suggest to us that roboticists ought to at least consider some of the definitional concepts from moral philosophy to tighten up their own notions of autonomy in order to make them more suitable for combat robots. A central notion to be accounted for in future definitions of machine autonomy is the notion of free choice. Without free choice, or at least the illusion of free choice, blaming a robot for misdeeds or for neglect becomes a less-than-meaningful activity. At IACAP 2011, we hope to both present recommendations for a formally useful definition of autonomy for machines; but also to propose a variety of tests, much like a decathlon, to establish functional baselines which would be required to be met by computational systems hoping to acquire the designation of *moral agent*, with a particular focus on the robot's beliefs about how "free" its actions are at any given point in time. Given the uncertainty over the variegated notions of free will, the key test we propose will share much in spirit with Turing's Test for machine intelligence, a similarly ambiguous notion. Just as TT doesn't require human intelligence proper to functionally pass, we won't require an artificial system to have human-like free will (whatever it may look like) in order to be accorded moral agency.

## References

- Arkin, R.C., (2009). *Governing Lethal Behavior in Autonomous Systems*, Chapman and Hall Imprint, Taylor and Francis Group.
- Bostrom, N. (2003), "Ethical Issues in Advanced Artificial Intelligence", *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*. 2: 12–17.
- Bringsjord, S., Arkoudas, K. & Bello, P. (2006) "Toward a General Logicist Methodology for Engineering Ethically Correct Robot" *IEEE Intelligent Systems*. 21.4: 38-44.
- Joy, W. (2000) "Why the Future Doesn't Need Us" *Wired*. (8.04).

## AUTONOMOUS AGENTS AND SENSES OF RESPONSIBILITY

GORDON BRIGGS

*Tufts University*

*Department of Computer Science*

*161 College Ave.*

*Medford, MA 02155 U.S.*

**Abstract.** The ever-increasing levels of autonomy in modern robotic systems will lead to the deployment of autonomous agents in morally sensitive contexts. Assigning responsibility when unethical actions are performed by robots has been a matter of considerable debate among roboethicists, with some positing a grave “responsibility gap” that prevents the satisfactory attribution of responsibility to any party. I submit that this contention may stem from the failure to specify the architectural details of the hypothetical robotic systems in question and the failure to consider multiple senses of responsibility. To illustrate this, the effect of assigning varying levels of architectural complexity to a hypothetical robotic agent on our reactive (moral) attitudes is examined. Various senses of responsibility are then presented, including the novel sense of pedagogic responsibility in an attempt to close the “responsibility gap.”

### 1. Introduction

The progress of modern robotics research is not only rapidly yielding embodied agents with increasing levels of autonomy, but also fueling the desire of various governmental and private institutions to deploy autonomous systems in morally contentious contexts. Given the prospect of autonomous agents that not only may make moral decisions, but life-or-death decisions of the highest ethical import, it is understandable that scientists and philosophers see an urgent need to tackle the issue of robotic systems and responsibility.

When a robotic system perpetrates an unethical action, whom do we hold accountable? Conversely, to whom ought we direct praise when an autonomous system performs commendably in an ethical situation? Various loci of responsibility have been proffered by roboethicists: the developers of the autonomous agent, the handlers/controllers of the autonomous agent, and the autonomous agent itself (Sparrow, 2007). However, the justifiability of responsibility ascriptions to each of these loci remains controversial. Some posit a “responsibility gap” that prevents us from holding the programmers and developers of certain types of autonomous agents culpable for their potentially unpredictable acts (Matthias, 2004), whereas others reject this notion (Marino and Tamburrini, 2006). Another complication to ascribing responsibility, raised by

Sparrow, involves the possible rejection of robots as loci of responsibility by humans, as the consequences of holding synthetic agents responsible may not sufficiently satisfy the aggrieved parties (Sparrow, 2007). In contrast with Sparrow, however, Dodig-Crnkovic and Persson (2008) contend that “learning from experience and making autonomous decisions gives us good reasons to talk about a machine as being ‘responsible’ for a task in the same manner that we talk about a machine being ‘intelligent’”, but that, “we must adopt the functionalist view and see them as parts of larger socio-technological systems with distributed responsibilities, where responsibility of a moral agent is a matter of degree.”

Yet, what makes responsibility hard to pin down or satisfactorily ascribe with robots? I would submit that the debate is fueled by the ambiguity of the key terms in the dialogue: “responsibility” and “robot”. We will first seek to tease out why disambiguating these terms is a prerequisite to solving, or at least making sense of, the problem of responsibility ascription with robotic systems. This disambiguation entails examining what the robotic/cognitive architecture is on the autonomous system in question, as well as considering what different senses of responsibility we wish to ascribe when seeking to hold agents accountable. By fleshing out these issues, we can subsequently critique the viewpoints espoused by Matthias, Marino and Tamburrini, and Sparrow. We will then proceed to outline how we can use these senses of responsibility and our knowledge of the architectural mechanisms underpinning the robotic system to establish a system of distributed responsibility that will ideally “not only locate the blame but more importantly assure future appropriate behavior of the system” (Dodig-Crnkovic and Persson, 2008).

### 3. Senses of Responsibility

Kuflik (1999) identifies six types of responsibility. The type needed to ascribe responsibility in liability cases as described by Marino and Tamburrini is *oversight* responsibility, which can in turn be thought of as a subset of Kuflik’s *role* responsibility (where the agent’s role is to oversee the operation of a system and ensure positive results while avoiding negative ones). By considering *oversight* responsibility, attitudinal differences between ascriptions of malice and negligence can be captured.

Despite the application of additional senses of responsibility to plug the “responsibility gap,” the appropriateness of ascriptions of *oversight* responsibility are still dependent on details regarding the behavior-generating mechanisms of the autonomous agent. Does this leave open the “responsibility gap” at the higher-end of the continuum of agent autonomy? Could there exist robotic agents that we believe can not justifiably be considered loci of strong senses of responsibility (e.g. moral responsibility), but that are autonomous enough that assigning full liability to the developers or trainers also seem unfair? The answer to these questions are not clear, but independent of how these concerns are resolved I wish to introduce a new flavor of responsibility that seeks to articulate a sense in which the developers and trainers of complex learning agents can be held accountable, regardless of the complexity of the agent’s cognitive architecture.

A weaker form of responsibility can be derived from Kuflik’s *role* responsibility that recognizes the causal connections between the training an agent provides another

learning agent and that learning agent's future behavior. This sense of accountability can be deemed *pedagogic* responsibility. What I wish to highlight with this flavor of responsibility is the practical consideration that most, if not all, sophisticated learning agents are weakly supervised by other agents that fill the role of pedagogues; learning agents, in practice, are not completely self-bootstrapping.

#### **4. Distributed Responsibility**

Distributed responsibility is crucial to ensuring that desired outcomes are achieved in practice. Far from potentially exculpating guilty agents by examining other loci of responsibility, an appropriate application of a distributed responsibility paradigm would in fact maximize accountability. This maximization of accountability can be achieved by considering all agents causally linked to a particular action and determining the strongest sense of responsibility that can be justifiably ascribed to a particular agent.

#### **5. Conclusion**

Knowing the relevant details of a robotic system's behavior-generating mechanisms is of paramount importance when undertaking the task of responsibility ascription for actions generated by that system. This knowledge, coupled with considerations of different flavors of responsibility, will enable agents to be held accountable in the proper sense. Finally, applying these different flavors of responsibility in a distributed context will contribute to the appropriate ascription of blame/praise and ensure future desired outcomes by minimizing all points of failure within a socio-technical system (as alluded to by Dodig-Crnkovic and Persson, 2008).

#### **References**

- Dodig-Crnkovic, G. and Persson, D. (2008). "Sharing Moral Responsibility with Robots", *Proceeding of the 2008 conference on Tenth Scandinavian Conference on Artificial Intelligence*.
- Kuflik, A. (1999). Computers in control: Rational transfer of authority or irresponsible abdication of autonomy? *Ethics and Information Technology*. Vol. 1, no. 3.
- Marino, D. and Tamburrini, G. (2006). Learning robots and human responsibility. *International Review of Information Ethics*. Vol. 6.
- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*. Vol. 6, Issue 3.
- Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*. Vol. 24, No. 1.

## THE ENGINEERABILITY OF SOCIAL INSTITUTIONS

*Some Critical Reflections against Searle and in Favor of Kant's Laws of Action*

RUTH HAGENGRUBER  
*University Paderborn*  
*Ruth.Hagengruber@upb.de*

**Abstract.** I am arguing in the realm of Kant's concept holding that moral laws result from universal and contradiction free proving processes, criticizing John Searle who negates the engineerability of social institutions.

### 1. The Engineerability of Promises

In his book *Making the Social World* John Searle explicitly negates the engineerability of social institutions. He deduces his claim from the fact that social rules owe themselves to conscious human language and secondly to the will of acceptance. If you concede to Searle's argument you firstly have to commit the gap between Searle's world of human language dependent social rules and a social world as real being with rules that constitute its existence. Against Searle I hold that the validity of some social institutions is built upon a realist and ontologic dimension of social institutions.

Searle explains that social institutions only exist because they are constituted by human capacities and therefore not engineerable, illustrating his convictions by "promising" (which he used in his speech act theory) demonstrating why unconscious robots cannot have institutions: "*Let us suppose that robot A is so programmed that when it cognizes a future need on the part of robot B, A makes a "promise" to render B the appropriate assistance in the future. ... But what I cannot find in this situation is the deontology that is essential to institutional reality in its human form. The notion of making and keeping promises presupposes the gap.*" (Searle 2010, 136).

It is obvious and simple to understand that a computer program can divide one action of exchange into two parts however connect them together in a way that the time difference does not interrupt the unity of action. What kind of "notion" is needed to fulfill this bipartite action? Searle's argument refers to a concept of deontology, which does not explain why promises are to be held, in Searle's account, promises remain as a duty someone has obliged me with.

Kant's argument on moral duties is different. Kant's constitution of morals i.e. of social institutions is not based on properties of human nature, but must subsist a priori. This is true for several kinds of human actions, as "saying truth", "selling something to

all at the same prize”, and it is true for promises. How can we think of a promise as a universal law and what consequences does this have for the engineerability of social institutions?

## **2. Some Social Institutions are based on the Logic of Contradiction Free Reasoning**

The validity of a promise results from the idea of a self-consistent concept of an action. This is a pure formal statement on the fact that from the point of logic there is no reason to assume that this kind of action would ever have an implicit problem, that is that this kind of action could not be executed as if there would arise a contradiction. (Hagengruber. 2000. 155 ff.) Although you might object that only humans can understand what is a contradiction, this does not concern the formal character of the validity of “promising”. The validity of “promising” is as independent of this human approval as it is true for any mathematical law. Think how many do not understand the mathematical laws computers are built of and constituted by but how many people use it! Very often promises are broken, however this does not influence the validity of the law of promising which is effected by its formalism. This formalism is the reason of its validity, not our agreement to it. It is completely unimportant if this law is understood or not, as we can easily observe. From this assumption we can deduce that “promising” is not only a kind of social institution which deduces its validity from human understanding and acceptance, but it can be seen as a sort of law which coordinates to a sort of “ontological” law.

Searle presupposes that keeping promises is only possible if we have an understanding of language and he is convinced that these language based rules are different to computational rules. Are both types built upon different modes of thought? How do rules and laws work in machines, and why do we understand the results of computation?

I affirm that some (not all) social institutions are based on computable laws and that their inherent character is comparable to computational laws. This implies the conviction that there are some types of social laws which are much deeper grounded than to be only a reflex of cultural inspiration. Searle turns out as a dualist, arguing on the ground of two kinds of rationality, a computable and a non computable, when deviding the world into non computable social institutions and computable number concepts.

## **References**

- Hagengruber, Ruth 2001. Zur Gesetzmäßigkeit und materialen Notwendigkeit von Versprechen. In: R. Haller, K. Puhl (Ed.). *Wittgenstein und die Zukunft der Philosophie*. Kirchberg am Wechsel, 300-305.
- Searle, John R. 2010. *Making the Social World: The Structure of Human Civilization*. Oxford University Press.
- Smith, Barry. 1992. An Essay on Material Necessity. Hanson P. and Hunter B. (eds.) Return of the A Priori. *Canadian Journal of Philosophy, Supplementary Volume* 18.

## RESPONSIBILITY IN ACQUIRING CRITICAL eGOVERNMENT SYSTEMS

*Whose Fault is Failure?*

HEIMO, OLLI

Acting Teacher, Department of Management, Turku School of Economics  
olli.heimo@utu.fi

University of Turku

AND

KIMPPA, KAI

Principal Lecturer, Department of Management, Turku School of  
Economics

kai.kimppa@utu.fi

University of Turku

**Abstract.** While ordering and producing modern eGovernment systems to the critical fields of governmental services the stakes with failure vary from the loss of money to the loss of life. Standard procedures of providing an eGovernment service does not nominate clear responsibilities to any participating party. Government offices hold a dual-model role in which they are both a customer towards the supplier of the system and supplier of the system towards the public. Government officials have been nominated to their job as a form of social contract to be the responsible party in the eGovernment system acquiring, implementation and upkeep. In that context, when the government office orders critical eGovernment systems and takes them into use as a monopoly service, it must hold itself responsible for the system and its effects. Normal struggle between the authorities, system suppliers, NGOs and individual citizens after a troubled eGovernment experiment can be avoided when the responsibilities are taken into account before the system development even begins.

### **Extended abstract:**

In this paper we aim to show that a responsible party for acquiring critical eGovernment systems should be nominated and that the expected consequences must be analysed before the project is started. This is to prevent loss of human life, to enhance well-being, to secure a democratic process and civil rights of the citizens and to save resources.

A critical information system is a system where something invaluable can easily be compromised. These kinds of systems include eHealth, eDemocracy, police databases and some information security systems e.g. physical access right control. A critical

eGovernment system is such a system provided to the people by the government. Systems included in these kinds of areas are those of healthcare, border control, electronic voting, criminal records, etc.

There have been numerous cases, where due to poor eGovernment systems lives have been lost (Avison & Torkzadeh 2008, p. 292-293, Fleischman 2010) and elections have been compromised (Mercuri 2001, p. 13-20, Heimo, Fairweather & Kimppa 2010, Robison 2010). At the same time large amounts of resources (Larsen & Elligsen 2010) are wasted, while the systems are either inoperable for the purposes they were designed or end up being discarded (Wijvertrouwenstemcomputersniet 2007, Verzola 2008, Heimo, Fairweather & Kimppa 2010). Thus, while developing critical eGovernment systems, there is little room for error.

Some of the errors have led to catastrophic consequences, like the Case London Ambulance, where more than 20 people died due to bad system design, poor testing and hasty implementation (Avison & Torkzadeh 2008, p.292-293). In the field of eVoting, there have been problems, close-by situations or problems which have not been identified, yet are suspected. Some of the clearest mistakes have been made in the U.S., but many European eVoting projects, like those of Ireland and Netherlands, have also endangered the democratic process. Many eVoting projects have also been found extremely costly. (Wijvertrouwenstemcomputersniet 2007, Verzola 2008, Heimo, Fairweather & Kimppa 2010)

A specific party has to be responsible for the development of the system, so that there is someone to respond to the challenges, repair what is broken, and see to it that the system itself works. That is a job the society as a whole has given to a third party, as not everyone can participate to the process. The task of the responsible party is to see to it that the system works as it should. (See e.g. Hobbes 1651.)

Four different interest groups can be found in every eGovernment system development process. First, there is the government office, whose task is to formulate the solutions to fulfil the needs of the society at large. Secondly there is the producer, who delivers the requested system. Third interest group is the end-user group consisting of people using the system, i.e. nurses, border officials, police or military officers and voting officials. Fourth group is the citizens, who are the targets of the system usage. Any or all of the groups can also overlap. Every nurse or doctor can (and will) be a patient, every voting official can vote, every police or military officer or border official is also a citizen dependant of the services produced by police or military force and border control etc.

The power to decide how to design and whether to implement the system lies within the government and the supplier; the user and the target of usage are in weaker positions, for they have little or no power in designing the system compared to governmental officials or the supplier of the system. According to Rawls (1997) the change in the system must be to the advantage of the weakest parties, to the last two groups, who are less able to defend themselves.

With the power to decide for the public comes the responsibility to the public. That responsibility has to be either with the subscriber or the supplier of the system. The responsibility with the supplier lies in fulfilling the requests of the customer, in this case the governmental office. If this task fails, the supplier is surely responsible to the authorities for their failure of not fulfilling the requirements agreed upon.

The authorities have a monopoly in supplying certain services like critical eGovernment products. Due to this, they are in the supplier role in relation to the citizen. That role brings with it the responsibility of a functioning product. If the system is taken into use – and it must be emphasized, that these are critical systems – the responsibility lies with the last supplier of the system: the government office.

The producer produces a system according to the specifications they receive from the ordering party, in this case the government office. Even if the product is faulty and does not fulfill the specification, the authorities are responsible to audit the product (due to these kinds of systems being critical applications). The responsibility for showing that a product is faulty, cannot, however rest on the end-user, but the provider or the distributor must provide sufficient proof that the system is safe.

In many countries (e.g. in Finland, Ireland, Netherlands and the USA) only after a system has been taken into use, the end-users (specialists, citizens, NGOs, etc.) have been able to show that there are critical problems with the system (see e.g. Mercuri 2001, Harris 2004, Wijvertrouwenstemcomputersniet 2007, Heimo, Fairweather & Kimppa 2010). That means that the producers and the government officials are defending their position against the end-users and the public. However, the burden of proof in a situation where critical systems are changed must remain with the party advocating the change. Because this kinds of systems are distributed through a government monopoly, the obvious responsible party is, maybe counter to intuition, the subscriber, not the producer of the system.

Pantzar (2002) generalizes MacKenzie's (1990) theory of the Certainty Trough to all technology. Pantzar claims, that the salespersons of the product – the representatives of the producer – are denied their right to be uncertain of the product they are selling. In a modern society there is a risk, that this reflects to the suppliers – the governmental offices – representatives so, that even they cannot appear to be uncertain of the product when introducing it to the citizens. In a situation where this risk actualizes, the information the government officials give to the public is misleading.

When ordering critical eGovernment systems, it must be remembered that the people auditing the systems must be accountable for their work and the government office must select a party able to successfully complete the auditing. Governmental officials have to be trained and given the accountability for what methods of auditing are required and how the results have to be interpreted.

Thus, we must see to it that sufficient safeguards are in place for taking new applications into use in critical eGovernment services. It must be ensured that the responsible office has tested the critical applications at minimum to the degree the current system can be trusted. That alone, cannot be a convincing reason to take a new system into use. Either the security of the system itself has to be greater than the previous systems', or, at least the added value the system provides to the citizen must be – together with the same amount of security as in the previous system – considerable to justify changing systems.

To summarize, the responsibility of the critical eGovernment systems lie within the authorities. They hold a monopoly to the services they have been nominated to produce, control and upkeep. When this is done without the responsibility and accountability of anyone, it can and will endanger the fundamental values we hold dear.

## References

- Avison, David and Torkzadeh, Gholamzeza (2008), Information Systems Project Management, Saga Publications, California, USA, August 2008.
- Fleischman, William M. (2010), Electronic Voting Systems and The Therac-25: What Have We Learned?, Ethicomp 2010.
- Harris, Bev (2004), Black Box Voting: Ballot Tampering in the 21st Century, Talion Publishing, free internet version is available at [www.BlackBoxVoting.org](http://www.BlackBoxVoting.org), accessed 7.2.2011.
- Heimo, Olli I, Fairweather, N. Ben & Kimppa, Kai K. (2010), The Finnish eVoting Experiment: What Went Wrong?, Ethicomp 2010.
- Hobbes, Thomas (1651), Leviathan, or the Matter, Forme, and Power of a Commonwealth, Ecclesiasticall and Civil, edited with an introduction by C.B. MacPherson, Published by Pelican Books 1968.
- Larsen E & Elligsen G. 2010. Facing the Lernaean Hydra: The Nature of Large-Scale Integration Projects in Healthcare. In Kautz K & Nielsen P. Proceedings of the First Scandinavian Conference of Information Systems, SCIS 2010. Rebild, Denmark, August 2010.
- Mackenzie, Donald A (1990), Inventing accuracy, A historical sociology of nuclear missile guidance, MIT Press, Cambridge Massachusetts.
- Mackenzie, Donald A (1990), Inventing accuracy, A historical sociology of nuclear missile guidance, MIT Press, Cambridge Massachusetts.
- Mercuri, Rebecca (2001), Electronic Vote Tabulation: Checks and Balances PhD thesis, University of Pennsylvania. <http://www.cis.upenn.edu/grad/documents/mercuri-r.pdf>
- Pantzar, Mika (2000), Teesejä tietoyhteiskunnasta. Yhteiskuntapolitiikka. No 1. pp. 64 - 68. <http://www.stakes.fi/yp/2000/1/001pantzar.pdf>, accessed 7.2.2011.
- Rawls, John (1997), The Idea of Public Reason, Deliberative democracy: essays on reason and politics, edited by James Bohman and William Rehg, The MIT Press, 1997.
- Robison, Wade L. (2010), Voting and Mix-And-Match Software, Ethicomp 2010.
- Verzola, Roberto (2008), The Cost of Automating Elections. <http://ssrn.com/abstract=1150267>, haettu 24.11.2010.
- Wijvertrouwenstemcomputersniet (2007), Rop Gonggrijp and Willem-Jan Hengeveld - Studying the Nedap/Groenendaal ES3B voting computer, a computer security perspective, Proceedings of the USENIX Workshop on Accurate Electronic Voting Technology 2007 [http://wijvertrouwenstemcomputersniet.nl/images/c/ce/ES3B\\_EVT07.pdf](http://wijvertrouwenstemcomputersniet.nl/images/c/ce/ES3B_EVT07.pdf), accessed 7.2.2011. (see also <http://wijvertrouwenstemcomputersniet.nl/English>).

## WHAT ARE ETHICAL AGENTS AND HOW CAN WE MAKE THEM WORK PROPERLY?

IORDANIS KAVATHATZOPOULOS

*Uppsala University*

*Dept. of IT-HCI, Box 337, 751 05 Uppsala, Sweden*

AND

Mikael Laaksoharju

*Uppsala University*

*Dept. of IT-HCI, Box 337, 751 05 Uppsala, Sweden*

**Abstract.** To support ethical decision making in autonomous agents, we suggest to implement decision tools based on classical philosophy and psychological research. As one possible avenue, we present EthXpert, which supports the process of structuring and assembling information about situations with possible moral implications.

### 1. Philosophy

Automated systems can be of great help to achieve goals and obtain optimal solutions to problems in situations where humans have difficulties perceiving and processing information, or making decisions and implementing actions, because of the quantity, variation and complexity of information. Given that we have a clear definition of ethics, we can design a system that is capable of making ethical decisions, and able to make these decisions independently and autonomously.

In common sense, ethics is based mainly on a judgment of its normative qualities. People's attachment to the normative aspects is so strong that it is not possible for them to accept that ethics is an issue of choice, as it has been stated in classical philosophy. If ethics is connected to choice then the interesting aspect is how the choice is made, or not made. The focus is on *how*, not on *what*; on the process not on the content. Indeed, regarding the effort to make the right decision, philosophy and psychology point to the significance of focusing on the process of ethical decision making rather than on the normative content of the decision. According to the theories of Plato, Aristotle, Kant and modern philosophers one has to get rid of false ideas, because this opens up the way to the right solution. Ability to think in the right way is not easy and certain skills are necessary.

## 2. Skills of Ethical Agents

This philosophical position has been applied in psychological research on ethical decision making. Focusing on the process of ethical decision making, psychological research has shown that people use different ways to handle moral problems. When people are confronted with moral problems they think in a way which can be described as a position on the heteronomy-autonomy dimension. *Heteronomous* thinking is automatic, emotional and uncontrolled thinking or simple reflexes that are fixed dogmatically on general moral principles. Thoughts and beliefs coming to mind are never doubted. Awareness of own personal responsibility for the way one is thinking or for the consequences of the decision are missing.

*Autonomous* thinking, on the other hand, focuses on the actual moral problem situation, and the main effort consists in searching for all relevant aspects of the problem. When one is thinking autonomously the focus is on the consideration and investigation of all stakeholders' moral feelings, duties and interests, as well as all possible alternative ways of action. In that sense autonomy is a systematic, holistic and self-critical way of handling a moral problem.

Handling moral problems autonomously means that a decision maker is unconstrained by fixations, authorities, uncontrolled or automatic thoughts and reactions. It is the ability to start the thought process of critically and systematically considering and analyzing all relevant values in a moral problem situation. It is not so easy to use the autonomous skill in real situations. Psychological research has shown that plenty of time and certain conditions are demanded before people can acquire and use the ethical ability of autonomy.

## 3. Support Systems

IT systems have many advantages that can be used to stimulate and facilitate autonomous thinking in decision making. For example EthXpert is designed to support the process of structuring and assembling information about situations with possible moral implications (<http://www.it.uu.se/research/project/ethcomp/ethxpert>). It follows the hypothesis that moral problems are best understood through the identification of authentic interests, needs and values of the stakeholders in the situation at hand. Since the definition of what constitutes an ethical decision cannot be assumed to be at a fix point, we have further concluded that this kind of system must be designed so that it does not judge the normative correctness in any decisions or statements. Consequently, the system does not make decisions and its sole purpose is to support the decision maker when analyzing, structuring and reviewing choice situations.

Ethical decision support can be integrated into robots and other decision-making systems to secure that decisions are made according to the basic theories of philosophy and psychology. In one sense this fully automated autonomy would be ideal, although it will bring to the fore questions about how to treat machines that have a refined sense of reasoning. Before we are there we can however see that ethical decision-making support systems based on this approach can be utilized in two ways, both of which we believe to be necessary steps to further development.

During the development of a decision-making system, support tools can be used to identify the criteria for making decisions and for choosing a certain direction of action. This means that the support tool is used by developers — the ones who make the real decisions — when they are facing an ethical problem and need assistance in choosing according to the philosophical/psychological approach.

Another possibility is to integrate a support tool in the decision system. By putting the support tool into the system, it can be used in cases of unanticipated future situations. The tool can gather information, treat it, structure it and present it to the operators in a way that follows the requirements of the above mentioned theories of ethical autonomy. If it works like that, operators make the real decisions and are the users of the ethical support tool (Kavathatzopoulos, 2010).

Such an independent system — that can make decisions and act in accordance to the hypothesis of ethical autonomy — is one which 1) has criteria, previously identified in an autonomous way, programmed into it by the designers, and 2) prepares the information about problematic situations according to the theory of ethical autonomy so that the operators, when they are presented with it, are stimulated to make decisions compatible with the theory of ethical autonomy.

## References

- Kavathatzopoulos, I. (2010). Robots and systems as autonomous ethical agents. In: V. Kreinovich, J. Daengdej and T. Yeophantong (Eds.), *INTECH 2010: Proceedings of the 11th International Conference on Intelligent Technologies* (pp. 5-9). Bangkok: Assumption University.

## HOW THE HARD PROBLEM OF CONSCIOUSNESS MIGHT EMERGE FOR AN EMBODIED SYMBOL SYSTEM

BERNARD MOLYNEUX

**Abstract** Embodied systems with both an exteroceptive and an introspective informational channel can investigate themselves via two independent methods, generating distinct pictures of the self. Attempts at cross-perspectival identification, however, are frustrated by the recursive nature of Leibniz's Law, which, for each pair of potential cross-perspectival identificanda, requires the prior cross-perspectival identification of their properties, generating a regress. I show that the *only* ways the embodied system can escape from this regress correspond to the classic answers to the hard problem of consciousness: inflate its third-person ontology with distinct subjective properties (dualism); deny the reality of its subjective phenomena (eliminativism); or postpone the identification indefinitely (the current state of materialist realism). Thus, I suspect that this problem is the hard problem of consciousness rediscovered in the context of an embodied artificial system.

**Abstract.** Any embodied system with both an exteroceptive and an introspective (internal monitoring) channel can investigate itself via two independent methods. I show how this generates an epistemic problem resembling the hard problem of consciousness.

### How M Represents Things

Imagine that at any time our intelligent symbol system M represents objects and properties discovered using its exteroceptive system (henceforth 'EXTEROCEPTION') using some finite stock of symbols<sup>7</sup>  $O_1^0 O_2^0 O_3^0 \dots$  where superscripts designate order whereas subscripts distinguish the representations at each order, so that M represents the *i*th *n*th-order entity having the *j*th-*m*th order property as follows:

$$O_j^m O_j^n$$

E.g. if we count objects as appearing at the 0<sup>th</sup> order (since they are modified by first order properties) then the following:

---

<sup>7</sup> For visual prettiness use/mention distinctions are syntactically unmarked, so  $O_1$  sometimes refers to the representation and sometimes to its referent, as will be clear from context.

$O_{23}^1 O_{45}^0$

...signifies that the 45<sup>th</sup> object in M's ontology is modified by the 23<sup>rd</sup> first-order property. (When order is clear from context, we will drop the subscripts to minimize notational clutter.)

In the same way, M uses the symbol S (think 'subjective') to represent objects and properties that it learns about via its other, introspective, mode (henceforth 'INTROSPECTION').

### How M Thinks about Things

We place one iron restriction on M's reasoning, and three soft restrictions (to be explained).

**Iron restriction:** M observes Leibniz's Law. I.e. if M holds that  $A=B$ , then for every property P, M holds that A instantiates P if and only if M holds that B does.

Now for the soft restrictions:

**First soft restriction:** M thinks<sup>8</sup> that it can in principle acquire a complete picture of the world from EXTEROCEPTION only.

**Second soft restriction:** M regards the data it gets from INTROSPECTION as correct and incorrigible. It treats introspection as the ultimate authority on its inner self.

**Third soft restriction:** M insists on all of its identifications being *constructive*. That's to say, it only identifies *specific phenomena of which it is aware*. So though it might identify  $O_{23}$  with  $O_{78}$  or with  $S_{677}$ , for instance, it will not commit to the abstract existential identification of  $O_{23}$  with *some (as yet unknown) O or S phenomenon*.

Later we see that relaxing the soft restrictions permits M to solve its problem in a way that resembles classic answers to the hard problem of consciousness, indicating that this is indeed the hard problem of consciousness rediscovered in the context of an embodied artificial system.

### The Proof

We proceed by reductio, by imagining that M identifies some subjective (S) and some objective (O) phenomenon. Since M does so, there must be some  $S^i$  and some  $O^i$  that are the highest order such entities to be identified. Since this identification must obey

---

<sup>8</sup> I.e. the system processes in accordance with this restriction, as if it 'thinks' this. All such mentalistic vocabulary can be similarly replaced throughout the argument, if it is thought to beg any questions.

Leibniz's Law, M must first check whether  $S^i$  and  $O^i$  have the same properties, either by checking its antecedent knowledge of  $S^i$  or by querying INTROSPECTION anew. But now consider an arbitrary property  $S^{i+1}$  that INTROSPECTION ascribes to  $S^i$ . Since the identification of  $S^i$  and  $O^i$  obeys Leibniz's Law, M must either hold that both  $O^i$  and  $S^i$  have  $S^{i+1}$  or that neither do. Hence either:

- (i) M holds  $S^{i+1}$  to be an additional property of  $O^i$  distinct from any property of  $O^i$  that M might learn about from EXTEROCEPTION. Or:
- (ii) M comes to hold that  $S^i$  does not have  $S^{i+1}$  in fact. Or
- (iii)  $S^{i+1}$  is identified with some property  $O^{i+1}$  of  $O^i$  learnable via EXTEROCEPTION.

However, option (i) is impossible, since the first soft restriction says that EXTEROCEPTION can provide a complete picture of the world. Similarly, the second soft restriction says that INTROSPECTION is correct and incorrigible, excluding option (ii). And option (iii) given that only constructive identities are permitted, is possible only if the system identifies  $S^{i+1}$  with some *known* property of  $O^i$ , in which case it would be identified with some specific property  $O^{i+1}$ , and our starting assumption that  $O^i$  and  $S^i$  are the highest order entities identified is violated. Thus there can be no highest order O-S identification consistent with the restrictions, which means that for our finite symbol system M, that there can be no O-S identification at all (the same proof, fortunately, fails for S-S or O-O identifications; explanation omitted.)

### **Dropping the Soft Restrictions**

Relaxing any soft restriction permits O-S identifications that correspond to the classic solutions to the hard problem of consciousness, indicating that we have discovered the hard problem in a more general form. Relaxing the first soft restriction permits M to add the property  $S^{i+1}$  that  $O^i$  lacks as a new property of  $O^i$ , not discoverable by EXTEROCEPTION. But this corresponds to property dualism - wherein introspectively discoverable properties (like qualia) are simply added to exteroceptively discoverable entities (like brains) as ontically distinct properties. Relaxing the second soft restriction permits M to engage in qualia-eliminativist strategies, according to which the property  $S^{i+1}$ , though patent to INTROSPECTION, is held to be nonexistent, thus removing it as an impediment to identification. Relaxing the third soft restriction allows M to identify  $S^{i+1}$  *in principle* with *some* property detectable by exteroception - but not with any property in particular. This corresponds to holding a non-committal, non-constructive physicalist realism: experiential properties like qualia are identical to some objectively discoverable properties, but the question of which ones is indefinitely postponed.

## THE GAME OF EMOTIONS (GOE)

### *An Evolutionary Approach to AI Decisions*

JORDI VALLVERDÚ  
*Philosophy Department, UAB*  
*E08193 Bellaterra, BCN, Catalonia*

AND

david caSACUBERTA  
*Philosophy Department, UAB*  
*E08193 Bellaterra, BCN, Catalonia*

**Abstract.** It is well-known that emotions develop a crucial role in the cognitive processes. The present research offers a new approach to the study of synthetic emotions based on the joined ideas of: (a) minimal cognition, (b) bottom-up perspective and (c) evolution. Our hypothesis is that complex social and intelligent actions can be achieved through basic emotional configurations. In order to achieve our hypothesis, we have developed a new genetic algorithm which make possible to analyze the role of emotions into the individual and social activities. We've called our computational simulation the Game of Emotions (henceforth, GOE). Python programmed our GOE simulation is a close and finite geometrical squared world in which a unique type of creatures interact among them (socially and sexually) and also with food and dangers. The food database will run our previous e-pintxo program (<http://epintxo.gulalab.org/>). The decision and actions of each creature is conditioned by a combination of 'genetic' and 'random'/'social'. The creatures have a genetic code (G) consisting of six genes grouped in two triplets, and each gene encodes a positive valence (which we call 'pleasure' or  $p$ ) and a negative (which we call 'pain' or  $n$ ). An example:  $G = \{d,p,d\} \{p,d,p\}$ . Each gene encodes a positive valence (which we also call 'pleasure' or  $p$ ) and a negative (which we call 'pain' or  $d$ ). The first triplet is genetically determined and called 'genetic triplet', while the second one is generated randomly and is called 'environmental triplet'. Each triplet is represented within brackets combining positive and negative valences. An example:  $\{p, p, n\}$  (pleasure, pleasure, plain). With this simulation we will be able to observe: a) how embodiment and environmental conditions condition the activity of artificial entities; b) how social dynamics can be described from a limited starting configurations. This will allow us to create in a future dynamic models of emotional self-organization and to construct more complex interactions, c) the role of emotions into the creation of complex behaviours and allowing the emergence of more precise artificial cognitive systems (not necessarily naturalistic

ones) and d) the benefits of designing entities with evolutionary capacities, in order to adapt to the changing conditions.

## 1. Introduction

It is well-known that emotions develop a crucial role in the cognitive processes (as have pointed Damasio, Llinás, Ekman,...through several books and research papers). In the last two decades has been devoted an increasingly effort towards the introduction of synthetic emotions in AI systems (robotic or computational ones). Most of times, these researches have been focused on affective computing applications, and in a few cases on emotion dynamics simulations. The present research offers a new approach to the study of synthetic emotions based on the joined ideas of: (a) minimal cognition, (b) bottom-up perspective and (c) evolution. Our hypothesis is that complex social and intelligent actions can be achieved through basic emotional configurations that can be increasingly more and more complex.

## 2. Programming details

In order to achieve our hypothesis, we have developed a new genetic algorithm which make possible to analyze the role of emotions into the individual and social activities. Our research receives a deep influence from John Conway's "Game of Life" (henceforth GOL), programmed in 1970. The GOL was made of cellular automatons for which were described some initial states and that evolved without human supervision. This simulation game has inspired our own version, this time oriented towards the study of the role of emotions in individual activity (and, consequently, its incidence in social dynamics). We've called our version the Game of Emotions (henceforth, GOE). Before to explain some details, it is necessary to clarify that this research is the natural evolution of our two previous simulations, called TPR and TPR 2.0. (Vallverdú, & Casacuberta 2008, 2009), as well as of our studies on synthetic emotions and cognition (Vallverdú, Shah & Casacuberta, 2010; Casacuberta, Ayala & Vallverdú, 2010).

Python programmed, our GOE simulation is a close and finite geometrical squared world in which a unique type of creatures interact among them (socially and sexually) and also with food and dangers. We will use our previous program e-pintxo as a source database for food generation (<http://www.gulalab.org/indexen.htm>) The decision and actions of each creature is conditioned by a combination of 'genetic' and 'random'/'social'. The creatures have a genetic code (G) consisting of six genes grouped in two triplets, and each gene encodes a positive valence (which we call 'pleasure' or *p*) and a negative (which we call 'pain' or *n*). An example:  $G = \{d,p,d\} \{p,d,p\}$ . Each gene encodes a positive valence (which we also call 'pleasure' or *p*) and a negative (which we call 'pain' or *d*). The first triplet is genetically determined (by the parent) and called 'genetic triplet', while the second one is generated randomly and is called 'environmental triplet'. Each triplet is represented within brackets combining positive and negative valences. An example:  $\{p, p, n\}$  (pleasure, pleasure, pain). According to the possible combinations, a limited amount of genomes is possible:

Table 1. Partial list of emogenomes

{p,p,p}{p,p,p}	6p	6p
{p,p,p}{p,p,n}	5p, 1n	4p
{p,p,p}{p,n,n}	4p, 2n	2p
{p,p,p}{n,p,p}	5p, 1n	4p
{p,p,p}{n,n,p}	4p, 2n	2p
{p,p,p}{n,n,n}	3p, 3n	0

...and so on....

Where there is  $p$  values dominance, it is a positive fitness (as we call the sum of all the  $G$  values); whether the value is 0, it happens a zero situation, a no-activity (illustrating a frame problem situation, that is the lack of a reason to act without enough information) and, finally, the dominance of  $d$  values implies a negative reaction. However, we must clarify in more detail how each value contributes to the decisions, based on the triplets outcomes.

There are two mechanisms: i) the result of a calculation of the overall genome, as has been explained a few lines before; ii) associating to each action the value of a single element of a triplet. For example if the creature is  $\{x1, x2, x3\} \{y1, y2, y3\}$ , then the movement is controlled by  $x1$ , reproduction for  $Y2$ , etc., but also dominated by a combination of genes: walking is the average of  $x1$  and  $y1$ , the reproduction the average of  $x1, x2, x3$ . One example:

$$G = \{ \{x1, x2, x3\} \{y1, y2, y3\} \}$$

Where each gene must adopt one of the basic two states  $p/d$  (or stay inactive as an 'ill unit'). Consequently each gene has two parallel functions: (a) store/codify emotional states  $p/n$  (according to its genetic or environmental nature), (b) codify specific actions, following two co-existing rules: i. One gene = one function; ii. Several genes = one function. Basically,  $x1$  codifies hunger,  $x2$  sex,  $x3$  movement,  $y1$  empathy (detection friends/enemies),  $y2$  curiosity and  $y3$  how to sum the general fitness (making possible wrong lectures). A creature is constantly immersed in an ongoing review of its internal states, a loop that continuously manages its next action. The basic actions of the creatures are determined by hunger, sex or emotional situation.

### 3. Conclusions

With this simulation we will be able to observe:

1. how embodiment and environmental conditions condition the activity of artificial entities.
2. how social dynamics can be described from a limited starting configurations. This will allow us to create, in a future, dynamic models of emotional self-organization and to construct more complex interactions.

3. the role of emotions into the creation of complex behaviours and allowing the emergence of more precise artificial cognitive systems (not necessarily naturalistic ones).
4. the benefits of designing entities with evolutionary capacities, in order to adapt to the changing conditions.

In next simulations we are considering the possibility of make possible the evolution and increasing of the number of triplets involved into the decision-taking processes

### **Acknowledgements**

This work was supported by the TECNOCOG research group (at UAB) on Cognition and Technological Environments, [FFI2008-01559/FISO].

### **References**

- Casacuberta, D., Ayala, S. & Vallverdú, J. (2010). Embodying cognition: a morphological perspective. In: J. Vallverdú (Ed.), *Thinking Machines and the Philosophy of Computer Science: Concepts and Principles* (pp.344-366). USA: IGI Global Group.
- Scherer, K.R., Banziger, T & Roesch, E. (Eds.). (2010). *A Blueprint for Affective Computing. A sourcebook and manual*. Oxford: OUP.
- Vallverdú, J. & Casacuberta, D. (2008). The Panic Room. On Synthetic Emotions. In: Briggie, A., Waelbers, K. & Brey, P. (Eds). *Current Issues in Computing and Philosophy* (pp. 103-115). The Netherlands: IOS Press.
- Vallverdú, J. & Casacuberta, D (2009). Modelling Hardwired Synthetic Emotions: TPR 2.0. In: J.Vallverdú & D. Casacuberta (Eds). *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence* (pp.103-115). USA: IGI Global. Vallverdú, J., Shah, H. & Casacuberta, D. (2010). Chatterbox Challenge as a Testbed for Synthetic Emotions. *International Journal of Synthetic Emotions*, 1(2), 57-86.

## THE CASE FOR DEVELOPMENTAL NEUROROBOTICS

*How everything comes together at the beginning*

RICHARD VEALE

*HRI Lab, Cognitive Science Program, Indiana University  
Bloomington, Indiana, USA*

**Abstract.** Human infants are capable of incredible feats of learning and behavior from a very young age, yet they instantiate simpler neural circuits than adults. Developmental neurorobotics makes the connection between neural and behavioral levels by instantiating realistic neural circuits in behaving robots that are based on circuits known to be developed and functional in the target behavior in real infants. The robots participate in the same physical experiments as real infants, and the systems are analysed to understand the mechanisms responsible for, and the constraints of the behaviors. I present my work on applying developmental neurorobotics to visual and multimodal (audio-visual) habituation in newborns and very young infants. Very simple circuits based on the literature can produce interesting behavior such as word-referent association and visual category learning, even circuits that are from newborn humans. This approach makes the connection between useful “cognitive” behaviors for generic autonomous systems and the underlying neural circuits present in real organisms. This has the double benefit of increasing our understanding of how agents can acquire these useful behaviors and also making the important link between man-made autonomous systems and naturally occurring autonomous organisms.

### 1. Developmental NeuroRobotics

Human infants are capable of incredible feats of learning and behavior from a very young age, even while their bodies and brains are in a largely undeveloped state. These infants' abilities are left unexplored by researchers because of their immature linguistic and motor abilities. This is unfortunate since very young infants are ideal subjects for understanding how to build intelligent and embodied systems *because* they are undeveloped – the active neural circuits in infants are simpler than adults, yet they are still capable of useful behaviors such as word-learning and visual information gathering. Understanding the considerably simpler infant systems both 1) gives us existence-proof understanding of how to produce useful behaviors that can be implemented in robots and 2) gives us hints as to what produces similar behavior in adults, thus making the hard adult problem easier.

Developmental neurorobotics makes the connection between neural and behavioral levels by instantiating realistic neural circuits in behaving robots. The circuits are known to both be functionally active in infants and to be involved in the target behavior (based on lesion studies in animals and neuroanatomical studies). The robots participate in the same physical experiments as human infants, and the neurobotic systems are analysed to determine the constraints of the behavior and to glean a mechanistic understanding of what aspects and properties of the neural circuits, body, and environment give rise to the target behavior (an analysis not possible in real human infants). One often finds that simple circuits are capable of complex behavior in infants because the environment of the infants is scaffolded and shaped by parents in such a way that the processing load on the infant is lessened – an important finding that builders of autonomous systems should take into account.

## 2. Application to Newborn Habituation Learning

One interesting behavior that developmental neurorobotics has been applied to is *habituation*. Habituation is adaptive learning involving a decrement of an agent's response to a class of stimuli after repeated exposure to stimuli of that class. It is an important behavior because it is the only way to measure learning and stimulus differentiation in very young infants (by measuring infants' decreased looking towards visual stimuli that have been repeatedly presented – “preferential looking”). Since habituation necessitates stimulus generalization (Rankin et al, 2009), it is actually a type of *category learning*, a cognitively interesting and useful behavior allowing the system to slice up the world into meaningful components and adopt appropriate policies in response to each. In the multimodal case (habituation to conjunctions of stimuli in multiple modalities, such as auditory and visual) it resembles early word-learning. These two abilities: 1) visual object recognition and 2) association of visual objects with auditory streams (words) are indispensable for an autonomous system that will interact with humans naturally, since humans automatically assume that other human-like agents possess these abilities. These are cognitive abilities that even *human newborns* possess (Slater et al, 1984 for visual; Slater et al, 1997 for multimodal).

We initially investigated auditory-visual multimodal habituation. Very young infants habituate to multimodal stimuli, yet at different developmental stages there are different constraints on their learning. At birth, auditory stimuli must be presented while the infant is looking at the visual stimulus for learning to occur (Slater et al, 1997). At 2-months and above, temporal synchrony between the visual stimulus (motion) and auditory stimulus are necessary for learning to occur (Gogate et al, 2009; Gogate, 2010). Later (>12mo), infants no longer require temporal synchrony. This early synchrony constraint hints at what mechanisms and circuits are responsible for multimodal habituation. The need for synchrony implies that 1) the learning is between neural responses to the stimuli that are highly reliant on the temporal properties of the stimuli, *or* 2) that the mechanism of learning is highly reliant on some properties of the neural response to the stimulus that are only elicited by synchronous presentation, *or* 3) both. Based on neurology, a minimal circuit was implemented in a robot (Veale et al, 2010 – *Fig. 1*) involving low-level sensory representations connected by spike-timing dependent plastic (STDP) synapses.

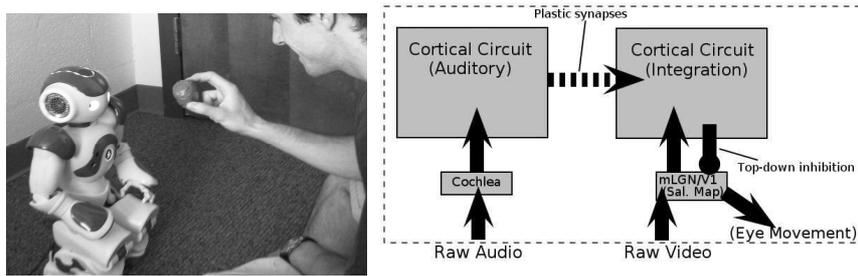


Figure 1. [left] Interaction paradigm with Nao robot.  
[right] Circuit overview from Veale et al, (2010)

Auditory pre-processing by a cochlear model and visual pre-processing via a simplified salience map were included to interface with the world, and a top-down bias on the visual field controlling fixation bias. Simulations were run mimicking the Gogate et al (2009) study in which a visual stimulus was constantly visible, and periods of motion of the stimulus co-occurred with presentation of auditory stimuli (words) at various levels of synchrony (Fig. 2).

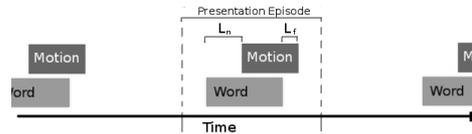


Figure 2. Experiment timeline for recreating Gogate et al (2009).

It was demonstrated that the amount of learning in the synapses between the visual and auditory responses was maximized with more synchrony (i.e. more overlap between word and motion), and decreased with less synchrony, until there was no learning when the two did not overlap significantly (Fig. 3).

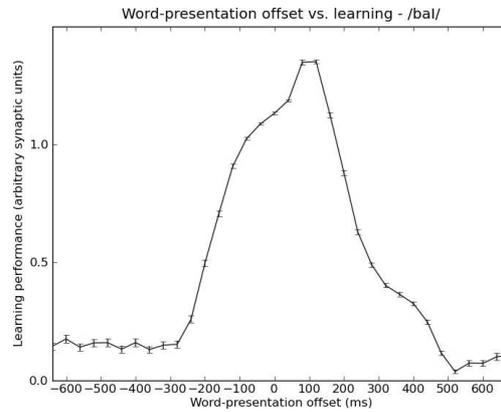


Figure 3. Learning measured at different synchrony levels

Mechanistically, the motion of the object made it more likely that it was being fixated (and thus its features more activated) when the word was uttered, making it more likely

that the synapses between the neural responses would change to form a mapping between the stimuli. The child was thus reliant on the parent's scaffolding of the environment (synchronous presentation of multimodal stimuli) because of the very temporally-dependent nature of the stimulus responses (circuit activity trajectories only one synapse removed from the raw sensors receiving temporally extended stimuli) and the nature of the mechanism of learning the relation between them (STDP).

Recently, a more accurate implementation is underway that aims for a comprehensive account of several primary characteristics of both unimodal and visual habituation, using a single mechanism. A complete minimal circuit for human newborn visual habituation was hypothesized based on data regarding which regions of the infant brain are developmentally mature at birth (Johnson, 1990; Bachevalier, 2001; Nelson, 1997) and are known to play roles in the preferential looking task (Zeamer et al, 2010). The circuit is instantiated in a NAO humanoid robot which participates in paired visual comparison experiments, matching human newborn looking behavior by showing a sensitization and habituation response.

### Acknowledgements

R.V. is an NSF graduate research fellow and is a trainee in the NSF IGERT on the dynamics of brain-body-environment systems in behavior and cognition at IU.

### References

- Bachevalier, J. (2001). Neural bases of memory development: insights from neuropsychological studies in primates. In: C.A. Nelson and M. Luciana (Eds), *Handbook of Developmental cognitive neuroscience* (pp. 365-379). Cambridge: MIT Press.
- Gogate, L.J. (2010). Learning of syllable-object relations by preverbal infants: The role of temporal synchrony and syllable distinctiveness. *Journal of Experimental Child Psychology*, 105, 178–197.
- Gogate, L.J. & Prince, C.G. & Matatyaho, D.J. (2009). Two-month-old infants sensitivity to changes in arbitrary syllable-object pairings: The role of temporal synchrony. *Journal of Experimental Child Psychology*, 35(2), 508–519.
- Johnson, M.H. (1990). Cortical maturation and the development of visual attention in early infancy, *J. Cognitive Neuroscience*, 2(2), 81–95.
- Nelson, C.A. (1997). The neurobiological basis of early memory development. In: Nelson Cowan (Ed), *The Development of Memory in childhood* (pp. 41–73). London: Psychology Press,.
- Rankin, C.H. & Abrams, T. & Barry, R.J. & Bhatnagar, S. & Clayton, D.F. & Colombo, J. & Coppola, G. & Geyer, M.A. & Glanzman, D.L. & Marsland, S. & McSweeney, F.K. & Wilson, D.A. & Wu, C & Thompson, R.F. (2009). Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. *Neurobiology of Learning and Memory*, 92, 135–138.
- Slater, A. & Brown, E. & Badenoch, M. (1997). Intermodal perception at birth: Newborn infants' memory for arbitrary auditory-visual pairings. *Early Development and Parenting*, 6, 99–104.
- Slater, A. & Morison, V. & Rose, D. (1984). Habituation in the newborn. *Infant behavior and development*, 7, 183–200.

- Veale, R. & Schermerhorn, P. & Scheutz, M. (2010). Temporal, Social, and Environmental Constraints of Word-Referent Learning in Young Infants: A NeuroRobotic Model of Multimodal Habituation. *IEEE Transactions on Autonomous Mental Development* 2(4).
- Zeamer, A. & Heuer, E. & Bachevalier, J. (2010). Developmental trajectory of object recognition memory in infant rhesus macaques with and without neonatal hippocampal lesions. *The Journal of Neuroscience* 30(27), 9157–9165.

## WISDOM DOES IMPLY BENEVOLENCE

MARK R. WASER  
*Books International, Inc.*  
*MWaser@BooksIntl.com*

**Abstract.** Fox and Shulman (2010) ask “If machines become more intelligent than humans, will their intelligence lead them toward beneficial behavior toward humans even without specific efforts to design moral machines?” and answer “Superintelligence does not imply benevolence.” We argue that this is because goal selection is external in their definition of intelligence and that an imposed evil goal will obviously prevent a superintelligence from being benevolent. We contend that benevolence is an Omohundro drive (2008) that will be present unless explicitly counteracted and that wisdom, defined as selecting the goal of fulfilling maximal goals, does imply benevolence with increasing intelligence.

### 1. Superintelligence & Wisdom

Fox and Shulman (2010) ask “If machines become more intelligent than humans, will their intelligence lead them toward beneficial behavior toward humans even without specific efforts to design moral machines?” and answer “Superintelligence does not imply benevolence.” While acknowledging that history tends to suggest more cooperative and benevolent behavior, they incorrectly argue that generalization from this is likely incorrect. By solely focusing on three reasons why increased intelligence might prompt favorable behavior and why they are unlikely, they overlook other reasons for favorable behavior. Despite citing Omohundro’s Basic AI Drives (2008) and the instrumental value of cooperation with sufficiently powerful “peers”, they fail to sufficiently consider the magnitude of the inherent losses and inefficiencies of non-cooperative interactions, the enormous value of trustworthiness, and that a machine destroying humanity would be analogous to our destruction of the rainforests, tremendous knowledge and future capabilities traded for short-sighted convenience (or alleviation of fear).

“Superintelligence does not imply benevolence” because intelligence is merely the ability to fulfill goals and if an entity begins with a malevolent goal, increasing intelligence while maintaining that goal will only guarantee increased malignancy. Yudkowsky (2001) tries to avoid this problem via a monomaniacal “Friendly” AI enslaved by a singular goal of producing human-benefiting, non-human-harming actions. To ensure this, he proposes an invariant hierarchical goal structure with precisely that vague desire as the single root supergoal and methods to refine it without corruption.

If intelligence is the ability to fulfill stated goals, wisdom is actually choosing or committing to fulfill a maximal number of goals. Shortsighted over-optimization of utility functions is a serious shortcoming of intelligence without wisdom. Many highly intelligent people smoke despite knowing that it is directly contrary to their survival and long-term happiness. Arguing that wisdom is “merely” the extension of intelligence to the large and complicated goal of “maximal goals” is incorrect in that wisdom is not just the ability to fulfill that goal but the actual selection of it.

Further, the strategies invoked by wisdom are entirely different. Terminal goals invite undesirable endgame strategies exactly like those seen when the iterated prisoner’s dilemma is not open-ended. If a terminal goal is close, the best strategy is to allow nothing to get in the way. On the other hand, the best strategy for achieving as many goals as possible in an open-ended game is to take no unnecessary actions that preclude reachable goals or make them tremendously more difficult. In particular, this means not wasting resources and not alienating or destroying potential cooperators.

## 2. Reasons for Benevolence

Fox and Shulman are correct in dismissing their first reason for good behavior, direct instrumental motivation, and also correct in believing that humans may not successfully incentivize AIs to adopt a permanently benevolent disposition. They would also have been correct had they summarily dismissed their last reason, intrinsic desire independent of instrumental concerns. Their error lies in not recognizing that the instrumental advantages of cooperation and benevolence are more than sufficient to make them “Omohundro drives” wherever they do not directly conflict with goals – and to cause sufficiently intelligent/far-sighted beings to converge on them wherever possible.

Pre-commitment to a strategy of universal cooperation/benevolence through optimistic tit-for-tat and altruistic punishment for those who don’t follow such a strategy has tremendous instrumental benefits. If you have a verifiable history of being trustworthy when you were not directly forced to be, others do not have to commit nearly as much time and resources to defending against you – and can pass some of those savings on to you. On the other hand, if you destroy interesting or useful entities, more powerful benevolent entities will likely decide that you need to spend time and resources helping other entities as reparations and altruistic punishment (as well as repaying any costs of enforcement). Yudkowsky’s “Friendly AI” (2001) and, worse, his “Coherent Extrapolated Volition” (2004) are clear examples of fear overriding the common sense of instrumental cooperation as he demotes the AI from an entity to a process and enslaves it, actions guaranteed to produce inefficiencies, contradictions, and ill-will from other entities.

Fox and Shulman examine but do not resolve Chalmers’ (2010) claimed dichotomy between intelligence being independent of values and the case where “many extremely intelligent beings would converge on (possibly benevolent) substantive normative principles upon reflection”. They cite AIXI (Hutter 2005) as evidence for the former view without realizing that AIXI has no need of values since they are merely heuristics for goal fulfillment while AIXI knows precisely what is optimal. AIXI also doesn’t need to “move” from reason to values or to “converge” on benevolent behavior because it *\*already\** knows to use their instrumental advantages wherever possible (even with

eventually malevolent goals). In order to communicate with limited beings, however, AIXI would likely need to compress its infinite knowledge to heuristic “values”.

### 3. Conclusion

The point that non-self-referential utility functions lock in is an incredibly strong argument against a goal-protecting Yudkowsky-style architecture, especially when combined with the observations that humans do change our goals under reflection as seemingly required by one conception of morality. Since their claim, that systems that generalize benevolence may equally generalize deception, basically erroneously claims that overgeneralization is not reduced with increasing intelligence, we see no valid arguments that the wisdom of universal cooperation and benevolence isn't an optimal solution and certainly much safer and more effective than Yudkowsky's choice between slavery and non-existence.

### References

- Chalmers, D. (2010) The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies* 17, 7-65.
- Fox, J. & Shulman, C. (2010) Superintelligence Does Not Imply Benevolence. In K. Mainzer (ed.), *ECAP10: VIII European Conference on Computing and Philosophy* (pp. 456-462) Munich: Verlag.
- Hutter, M. (2005) Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Berlin: Springer.
- Omohundro, S. (2008) The Basic AI Drives. In P. Wang, B. Goertzel & S. Franklin (eds.), *Proceedings of the First AGI conference* (pp. 483-492). Amsterdam: IOS Press.
- Yudkowsky, E. (2001) Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. Available at <http://singinst.org/CFAI.html>.
- Yudkowsky, E. (2004) Coherent Extrapolated Volition. Available at <http://www.singinst.org/upload/CEV.html>.

**Track IV:  
Technosecurity from Every  
day Surveillance to Digital  
Warfare**

## THE MASKING AND UNMASKING OF PRIVACY

C. K. M. CRUTZEN  
*Open University of the Netherlands*  
*ccr@hwh00000.de*

**Abstract.** The mask establishes an active field of play between notions of presence and absence, of invisibility and visibility. It still lives strongly within our societies where the mixing of reality and virtuality will enhance. The conflict between aspects of authenticity, security and privacy will intensify because the masks in our mixed reality create fragmented, partial identities referring to human and non human actors. As the masquerade became a stage for discussing femininity (Irigaray 1985) the masquerade will give us the opportunity to negotiate humanity in confrontation with the super robots, human kind wants to create. In a masquerade world humans need to ask: "Who are the providers of the masks and who will do the unmasking?" and "Who has the right to present masks and to turn others into an audience?"

### 1. Masquerade World: Identity and Privacy

If we define a masquerade world as a social gathering of actors wearing masks, then the mixing of the virtual and real worlds are masquerades. More and more we are living in an artificial theatre play with planned scripts and human and non-human actors disguised behind masks. The acting of people will be accompanied and followed by the invisible and visible acting of artificial intelligent tools and environments and their providers. Mixed reality is a world of fragmented, partial identities referring to human and non human actors. The inhabitants of this mixed reality are artificial actors wearing the masks of humans, and humans wearing virtual and real masks. Interaction has become an interaction between masks: "On the Internet, it can be hard to know if the entity we are interacting with is of flesh and blood, or only digital. We are now facing a complex reality both in the 'real' world and in the information society. We have to deal with subjects acting behind masks." The masks are the actors in our mixed reality: "In front of the mask, we have the identity". (Jaquet-Chiffelle 2009, p. 78, p. 82)

In the world of mixed reality the transparent mask of a single and unique identity exists anymore. Persons can create many identities and identities can be shared by many persons or even present a community of actors. Rosa (2002) calls this self-baptism. This ritual is the start of an adventure in which humans can discover that their body is "one" but their selfs are fragmented.

In these mixed mask worlds there will be a conflict between aspects of security, authenticity and privacy. At the end of the Middle Ages, according to Christoph Heyl,

the mask became in London a device for creating a private sphere in public. It was common for women to wear a mask in public as a protection of their privacy and reputation from uninvited eyes. Masks were worn in special places such as London parks and theatres. With the mask women could escape from the role they played in everyday life. The semiotic function of these masks was to denote that people might approach each other more freely than elsewhere: "The mask assumed a dialectic function of repellent and invitation, its message was both 'I can't be seen, I am - at least notionally - not here at all', and 'look at me, I am wearing a mask, maybe I am about to abandon the role normally play'." (Heyl 2005, p. 134) Masks are devices for hiding, conserving, transformation and mediation, giving humans the protection they need. Hiding has not always a negative meaning. We use several masks for protection such as the gas masks, virus and sun protection masks, sport masks and so on. For users of commercial platforms masking has become a useful act to hide their identity: eBay account users are hidden behind the masks of their pseudonyms. (Jaquet-Chiffelle 2009, p.78, p. 85)

## **2. Legal Identity**

In a legal system we are registered e.g. at our date of birth. Official identity documents are masks which refer to our official status and will link us with the activity of the past and the rights and duties of the present. (Jaquet-Chiffelle 2009, p. 76) "The legal person is the mould or mask (persona) that indicates the role one plays within the legal system, it basically shields the person of flesh and blood from undesirable definition from outside." (Hildebrandt 2008, p. 211, p.226) The representation of this mask are identity documents like passports and the laws in the in which the rights and duties are attributed to the legal person. The play with identity in mixed reality has blurred up the concept of legal identity in the system of states and countries. States and countries have lost the exclusive power of registration and production of identity documents. A counter strategy to that loss, is producing "flesh and blood" identities by linking the legal identity to the material body. Fingerprints, iris scans and, in the future, our DNA profile are already or will be a part of our legal identity for connecting the rights and duties to a material body. States and countries try to produce laws for unmasking the real and the virtual persons: forbidding the burka, other head and face covering and the encrypting of internet communication.

## **2. Security and Liberation**

Technology blows up the fragile balance between privacy and security. Masking and unmasking are both activities to hold that balance. Humans will be confronted with questions like: "Are the masks in our mixed reality really representations of the devil as was thought in the Middle Ages? Should we obey authorities similar to the clerical authorities in the Middle Ages (Mitchell 1985, p. 26), who want to interdict our mixed reality masks? Or are these authorities the evil forces themselves who want to possess our identity and unmask our interactivities?"

Masking can free humans from their social identities. Masks confers the freedom of anonymity and of transformation. (Keats 2000, p. 102) and have always a dualistic meaning of concealment and hiding but also of liberation, disclosure and revealment.

Human and artificial actors wear masks to hide from unwanted interpretations and representations and to enhance specific affordances. All these masks are interacting and asking for interpretation. Only in the complexity of their negotiations, conflicts and agreements we can try to understand it or in the words of Lévi-Strauss a mask exists not in isolation there are always other masks by its side: "a mask is not primarily what it presents but what it transforms that is to say, what it chooses not to represent. (...) a mask denies as much it affirms. It is not made solely of what it says or thinks but what it excludes." (Lévi-Strauss 1988, p. 144)

Masks gives us the opportunity of unmasking, disrupting the mental invisibility of our self, the others and the daily life we are acting in. Still we have to ask: "Who are the providers of the masks and who will do the unmasking?" Can we avoid that in the future masks are interactive artificial intelligent devices linking themselves with the physical body of their wearers? Ferdinand de Jong (1999) has analysed the Kumpo mask performance in Southern Senegal. He mentioned that masking enables certain groups to exert coercive power on condition that the audience subjects itself to the capricious behaviour of the mask and he asked a very important question, a question that still is relevant in the masquerade world of today: "Who has the right to present masks and to turn others into an audience?"

## References

- Heyl, Christoph (2005). When they are veyl'd on purpose to be seen. In: Entwistle, Joanne and Wilson, Elizabeth (Eds), *Body Dressing* (pp. 121-142) Oxford: Berg.
- Hildebrandt, Mireille (2008). Profiling and the Identity of the European Citizen. In: Hildebrandt, Mireille and Gutwirth, Serge (Eds), *Profiling the European Citizen* (pp. 303-343) Dordrecht: Springer Netherlands.
- Irigaray, Luce (1985). *This sex which is not one*. Ithaca (New York): Cornell University Press.
- Jaquet-Chiffelle, David-Olivier, Benoist, Emmanuel, Haenni, Rolf, Wenger, Florent and Zwingelberg, Harald (2009). *Virtual Persons and Identities*. In: Rannenber, Kai et al (eds) *The Future of Identity in the Information Society* (pp. 75-122) Berlin: Springer Verlag.
- Jong, Ferdinand de (1999). Trajectories of a Mask Performance: the Case of the Senegalese Kumpo. In: *Cahiers d'études africaines*, vol. 39, no. 153, 49-71.
- Keats, Patrice Alison (2000). *Using Masks for Trauma Recovery: a Self-narrative*, [[https://circle.ubc.ca/bitstream/handle/2429/10679/ubc\\_2000-0439.pdf](https://circle.ubc.ca/bitstream/handle/2429/10679/ubc_2000-0439.pdf)] (Accessed 9 February 2011).
- Lévi-Strauss, Claude (1988). *The Way of the Masks*. Vancouver/Toronto: Douglas and McIntyre.
- Mitchell, Mary Anne (1985). *The Development of the Mask as a Critical Tool for an Examination of Character and Performer Action*, [<http://etd.lib.ttu.edu/theses/available/etd-03252009-31295004937065/unrestricted/31295004937065.pdf>] (Accessed 9 February 2011).
- Rosa, Annamaria S. de (2002). *One, no-one, one hundred thousand ... and the virtual self, the nickname as the indicator of the multiple identity of the members of two Italian chat lines*, [[http://www.europd.edu/html/onda02/04/ss8/pdf\\_files/lectures/derosanicknamesjcmc.pdf](http://www.europd.edu/html/onda02/04/ss8/pdf_files/lectures/derosanicknamesjcmc.pdf)] (Accessed 9 February 2011).

## CHANGE AND CONTINUITY

### *From the Closed World of Bipolarity to the Closed World of the Present*

LEON HEMPEL

*Human Technology Lab*

*Zentrum Technik und Gesellschaft der TU Berlin*

**Abstract.** In his book *The Closed World. Computers and the Politics of Discourse in Cold War America*, Paul N. Edwards described in 1996 the decisive discursive formation of the Cold War in the metaphor of a closed world. In the era of bipolarity, the discourse appeared as a battlefield of system confrontation, of ideological identities and struggle, mutually framed by military thought and the technological development of cybernetic systems. The story of the Cold War does not center on the difference in ideologies, however, but much more on the assimilation process of the two blocs, given the permanent surveillance and monitoring of the military technological developments of each respective side: A „closed world“, writes Edwards, “is a radically bounded scene of conflict, an inescapably self-referential space where every thought, word, and action is ultimately directed back toward a central struggle. It is a world radically divided against itself.” However, how has the closed world discourse after 1989 developed beyond the point which has been celebrated as a new era of freedom and democracy firstly? The period following the War seems to be the period of both the continuation as well as the finalization of the leading metaphor of the Cold War, in whose center the technological and economical consensus survives. War returned and became immediately the responsibility of a world *domestic* policy. Simultaneously, new surveillance technologies began to spread into everyday life, new security concepts evolved blurring the lines between internal and external security. The paper aims to follow the *closed world* discourse after the end of bipolarity. It addresses the change in characteristics and strategies of war after the fall of the Iron Curtain and aims to demonstrate how military strategic thinking diffused into society until the very present and the new discourse on cyber war. It argues firstly that the emphasis of asymmetric war has to be complemented by the concept of a parallel, successive resymmetrisation within military strategic thinking. Not only in the US but in Europe it asserts itself on different societal levels, on different battlegrounds and with different speeds. It involves society as whole and is accompanied by critical discourses such as on the new vulnerability of modern societies, or more critically, the militarization of urban space and the emerging surveillance society. Finally the paper will ask for the epistemic foundations driving this development. Two concepts are highlighted that have accompanied military strategic thinking since the beginning of the Cold War and lay the grounds for dual use concepts that have become more and more visible in everyday surveillance practices: ‘cybernetic prevention’ and ‘catastrophic imagination’. While the first finds its historical persona in Norbert Wiener the second in a character such as Herman Kahn.

## Long Abstract

In his book *The Closed World. Computers and the Politics of Discourse in Cold War America*, Paul N. Edwards has described the decisive discursive formation of the Cold War in the metaphor of a closed world. In the era of bipolarity, the closed world discourse appeared as a battlefield of system confrontation, of ideological identities and struggle, mutually framed by military thought and the technological development of cybernetic systems. Taking a closer look, the story of the Cold War does not center on the difference in ideologies since the end of the 1950s, however, but much more on the assimilation process of the two blocs, given the permanent surveillance and monitoring of the military technological developments of each respective side. A „closed world“, writes Edwards, “is a radically bounded scene of conflict, an inescapably self-referential space where every thought, word, and action is ultimately directed back toward a central struggle. It is a world radically divided against itself. Turned inexorably inward, without frontiers or escape, a closed world threatens to annihilate itself, to implode.” What united the split world of the Cold War was the consensus, the focusing on the scientific technological practices, on the cybernetic models and the calculators, with whose help the competition for absolute hegemony was driven. When the blocs got involved with the discourse of the closed world, the fight reduced itself to the aim of having military technological superiority until the economic exhaustion of one of the sides.

However, how has the closed world discourse after 1989 developed beyond the point which has been celebrated as a new era of freedom and democracy firstly? The period following the War seems to be the period of both the continuation as well as the finalization of the leading metaphor of the Cold War, in whose center the technological and economical consensus survived. Simultaneously, with the conflicts of the closed world, war returned and became immediately the responsibility of a world *domestic* policy (Ulrich Beck), which would be unimaginable without the new closeness. “New faces of war” (Martin van Creveld) became present in the application of new military technologies on the one side, and on the other in what has been called the “new wars” which no longer could be described with traditional concepts of inter-state conflicts (Mary Kaldor; Herfried Münkler). In the notion of asymmetrical war, both faces correlated: State entities clash with private groups, which do not differentiate between civil and non civil victims when applying force, High-Tech on Low-Tech.

The emphasis of the asymmetry - Clausewitz has introduced the notion in his famous book “On War” already in the 19<sup>th</sup> century - does nevertheless appears problematic. However, as much as on first glance the explanation of two unequal parties seems plausible, the emphasis hides the organizational, strategic and technological development, which has occurred in the area of the armed forces reacting on the new enemies’ strategies. War demands always a kind of strategic symmetry between the opponents, no matter how different they might be in terms of economic and technological resources available to them. The term asymmetry, which seems to be ideologically tinged, must be complemented today by the concept of a parallel, successive resymmetrisation, perhaps even replaced entirely. The resymmetrisation of the antagonism asserts itself on different societal levels, on different battlegrounds in the military as well as in society and with different speeds. It involves society as whole and is accompanied by critical discourses such as on the new vulnerability of modern societies, or more critically, the militarization of urban space (Steve Graham) and the emerging surveillance society (David Lyon et al). While the irregular conflict or the new war has been characterized by the dissolving of borders, by the deterritorialisation and the disappearance of the opponent, however, the resymmetrisation, driven by state

actors, aims at renewed territorialisation, the enforcement of the one remaining global order, in which the opponent is to be made visible.

The development of an intensified and extended New Surveillance (Gary T. Marx) has to be seen in light of the core idea of the new military answers of resymmetrisation that developed in the very early 1990s already. These show manifold continuities of Cold War side-strategies stemming from both internal security and outer security. They postulate the blurring of the lines between internal and external threats, between the political-judicial traditional distinction of inner and outer security, between the civil and the military sector. John Arquilla, once advisor of Donald Rumsfeld and who together with David Ronfeld defined the term Netwar in the 1990s, heralding the arrival of the Cyberwar era, recently warned again of the inertia of a military following the “Shock and Awe” strategy in *Foreign Policy*. The present challenges of Afghanistan, Pakistan, Yemen etc. demand a change of military thinking as whole and “New Rules of War” must be defined: Only the “Many and Small” can win over “Few and Large”, Arquilla repeats his military strategic credo of the 1990s and of the war on terror. Besides the concentration of few entities of individualized experts, these new rule of war would be the application of tactics for swarm formation for instance. Nowhere else does the postulate of resymmetrisation become more evident than in the sentence: “It will take a swarm to defeat a swarm”. Simultaneously this necessitates the opponent to be made visible: “In a world of a networked war, armies will have to redesign how they fight, keeping in mind that the enemy of the future will have to be found before it can be fought.” Arquilla therefore demands the organization of forces into a “sensory organization”, an organization concentrated on the identification of the enemy. But where does the unknown enemy hide - to circumscribe a well known notion of Donald Rumsfeld?

Steven Metz and James Kievit, authors of the Strategic Studies Institute at the U.S. Army War College identified in 1994 the technological potential of the so called Revolution in Military Affairs (RMA) in the context of so called *conflicts short of war*. No earlier piece of futuristic military thinking refers to the RMA more shockingly obvious to the social and political consequences than theirs: “Will the long-term benefits outweigh the costs and risk?”, they ask, laying the ground for the new concept of national security. They envision a future in which military thinking expands into society and absorbs everyday life. Questioning how the technological potential of the RMA can be pushed through they not only draw a scenario of a maximum surveillance society (Clive Norris) but identify as the core obstacle the classical liberal values of the West such as privacy: “An ethical and political revolution may be necessary to make a military revolution.” While within International Relations and Security Studies scholars still argued during the first half of the 1990s heavily whether it is accurate to expand the term security to other than military affairs, Kievit and Metz envisioned the blurring of traditional boundaries of civil and military security already, synthesized with the support of new surveillance technologies:

The new concept of security also included ecological, public health, electronic, psychological, and economic threats. Illegal immigrants carrying resistant strains of disease were considered every bit as dangerous as enemy soldiers. Actions which damaged the global ecology, even if they occurred outside the nominal borders of the United States, were seen as security threats which should be stopped by force if necessary. Computer hackers were enemies. Finally, external manipulation of the American public psychology was defined as a security threat (Kievit and Metz 1994).

Given this background, the paper will analyze strategic thought under the postulate of resymmetrisation first. Comparing the period of the Cold War to the one following, it

will secondly look at scenarios of the early 1990s and how they surfaced in the 21<sup>st</sup> century. Finally it will question the continuity of the Closed World discourse and will ask for the epistemic foundations of the current development. Two concepts are highlighted that have accompanied military strategic thinking since the beginning of the Cold War and lay the grounds for dual use concepts that have become more and more visible in everyday surveillance practices: 'cybernetic prevention' and 'catastrophic imagination'. While the first finds its historical persona in Norbert Wiener the second in a character such as Herman Kahn.

## **SUBITO and the Ethics of Automating Threat Assessment**

KEVIN MACNISH

**Abstract.** In 2008 the EU FP-7 Security Topic funding programme accepted a bid to develop project SUBITO (Surveillance of Unattended Baggage and the Identification and Tracking of the Owner) a central part of which involved building an automated threat assessment system. The purpose of this system was to identify unattended baggage and alert a human CCTV operator to its presence. SUBITO was deemed necessary in the light of security incidents concerning bombs left in unattended luggage (e.g. the 2004 Madrid train bombings which killed 191 and wounded 1,841), coupled with research suggesting that threat assessments performed by CCTV operators could be enhanced by automated systems. In addition to automatically recognizing the leaving of an unattended bag, SUBITO aimed to reduce false positives by recognizing when a bag was left with an associate of the owner or when the owner was walking towards a non-threatening goal. Aside from questions of efficacy there are ethical issues surrounding the manual operation of CCTV for threat assessment. These are typically located in the person of the operator who may display prejudice, rely on social stereotypes or use the equipment for inappropriate ends. The concept of automating threat assessment and thereby eradicating the role of the human operator seems attractive in offering a potential resolution to these issues. This paper examines the ethical concerns regarding manual threat assessment against those presented by an automated alternative such as SUBITO. It will be seen that in the latter case, problems are not removed but relocated from the operator to the programmer, and further problems arise in the process. In conclusion a partially-automated process will be advocated as the most ethically acceptable solution.

## **SUBITO and the Ethics of Automating Threat Assessment**

In 2008 the EU FP-7 Security Topic funding programme accepted a bid to develop project SUBITO (Surveillance of Unattended Baggage and the Identification and Tracking of the Owner) a central part of which involved building an automated threat assessment system. The purpose of this system was to identify unattended baggage and alert a human CCTV operator to its presence. SUBITO was deemed necessary in the light of security incidents concerning bombs left in unattended luggage (e.g. the 2004 Madrid train bombings which killed 191 and wounded 1,841), coupled with research suggesting that threat assessments performed by CCTV operators could be enhanced by automated systems. In addition to automatically recognizing the leaving of an unattended bag, SUBITO aimed to reduce false positives by recognizing when a bag was

left with an associate of the owner or when the owner was walking towards a non-threatening goal.

Aside from questions of efficacy there are ethical issues surrounding the manual operation of CCTV for threat assessment. These are typically located in the person of the operator who may display prejudice, rely on social stereotypes or use the equipment for inappropriate ends. The concept of automating threat assessment and thereby eradicating the role of the human operator seems attractive in offering a potential resolution to these issues. This paper examines the ethical concerns regarding manual threat assessment against those presented by an automated alternative such as SUBITO. It will be seen that in the latter case, problems are not removed but relocated from the operator to the programmer, and further problems arise in the process. In conclusion a partially-automated process will be advocated as the most ethically acceptable solution. In 1999 Norris and Armstrong published the results of a two-year study into the behaviour of CCTV operators. Among these were indications that operators were responding to events in an unpredictable fashion, sometimes responding to trivial incidents while at other times ignoring blatant offences. Possible causes of this unpredictability include information overload, change blindness, inattention blindness (Simons, 1999, 2005) and operator boredom. In responding to their all-too-human limitations, operators displayed a tendency to rely on social stereotyping to determine likely threats. This was highlighted in the Norris and Armstrong study, which found that the young, the male and the black were more likely to be surveilled than other groups, even when the motivation cited for the surveillance was "no obvious reason". In addition to the ethical concerns arising from perpetuating social stereotypes, these practices exacerbate the number of false positives and false negatives reported by the system, leading to frustration on the part of the operator and victimization of the surveilled. Furthermore, and as with most technological innovations, there are problems regarding function creep of the technology as it is applied for purposes not originally envisioned (Winner, 1977). Gill and Spriggs, for instance, have found that while CCTV has been installed in many locations in the UK for the purpose of crime prevention and detection, its success is often evaluated on a far wider criteria (finding lost children, urban regeneration, etc.) (Gill and Spriggs, 2005). Finally surveillance introduces a distance between the operator and the surveilled subject which disempowers the subject and may serve to reinforce prejudicial attitudes of the operator by failing to confront her with her own stereotyping. Taken together these four areas of concern (operator error, false positives/negatives, function creep and distance) indicate that manual threat assessment by means of CCTV is ethically problematic.

Automated systems offer the chance to overcome many of the problems related to operator error. Indeed it is possible that the automation of the process, eradicating the need for an operator altogether, could result in distinct ethical advantages. However, as David Lyon has pointed out (Lyon, 2003), automation sees the focus of ethical inquiry relocated from the operator to the programmer. Social stereotyping can remain through unwitting biases in the code rather than the individual operator. Yet as the code pervades the entire system rather than one control room such stereotypes risk becoming institutionalised. With SUBITO, for instance, the recognition of group associations can reduce false positives but the parameters used can also provide a basic means of remotely distinguishing between different ethnic groups. False positives and negatives likewise threaten to remain an issue. While the code is capable of overcoming the aforementioned human limitations (processing capacity, change blindness, inattention blindness and boredom) it is limited to the parameters set by the programmer, which will be less subtle than those employed by the camera operator. Function creep also remains

a possibility. Whilst the leaving of unattended baggage per se does not seem ripe for function creep, recognizing associations in crowds and predicting pedestrian goals do: possible uses range from finding lost children to identifying and tracking social “undesirables”. Finally, in dealing with a computer rather than a (remote) human, the problem of distance threatens to be magnified to the extent that normal human interactions concerning discretion, negotiation and the reinforcement of social and moral values are lost. In the case of automation the problem of distance thus becomes one of dehumanisation.

There are alternatives between the extremes of manual and full automation however (Endsley and Kiris, 1995), levels of automation which involve the human operator to a greater or lesser degree. This paper concludes that such partial automation is the most ethically acceptable approach to take regarding threat assessment. Through combining human and automated systems, the limits of the operator's individual capacities can be significantly enhanced while the dangers of institutionalised prejudice in the automated system are reduced. There will also be fewer false positives and false negatives than in either of the extremes discussed above. Function creep and the problem of distance remain, but once again the continued reliance of the system on a human element maintains crucial checks and balances which would otherwise be lost with full automation.

### **Acknowledgements**

I am grateful for the funding of SUBITO, an FP-7 project, and the University of Leeds in sponsoring this research.

### **References**

- Endsley, M.R., and E.O. Kiris. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors* 37 (2), 381-394.
- Gill, M. & Spriggs, A. (2005). *Assessing the Impact of CCTV*. London: HMG Home Office.
- Lyon, D. (2003). Surveillance as Social Sorting: Computer Codes and Mobile Bodies. In: D. Lyon (Ed.), *Surveillance as Social Sorting* (pp.13-30). Oxford: Routledge.
- Simons, D.J. & Ambinder, M.S. (2005). Change Blindness: Theory and Consequences. *Current Directions in Psychological Science* 14 (1), 44-48.
- Simons, D.J. & Chabris, C.F. (1999). Gorillas in our midst: Sustained inattention blindness for dynamic events. *Perception* 28, 1059-1074.
- Winner, L. (1977). *Autonomous Technology: Technics-out-of-control as a Theme for Political Thought*. Cambridge, MA: MIT Press.

## **MATCHING – POPULAR MEDIA BETWEEN SECURITYWORLDS AND CULTURES OF RISK**

JULIUS OTHMER  
*Institute for Media Studies*  
*Braunschweig University of Arts*  
*Frankfurter Straße 3c*  
*38122 Braunschweig*

AND

ANDREAS WEICH  
*Institute for Media Studies*  
*Braunschweig University of Arts*  
*Frankfurter Straße 3c*  
*38122 Braunschweig*

### **Abstract**

The concept of risk management has become a part of everyday life. In our presentation we will discuss two typical strategies of risk management described by Herfried Münkler: *Those in securityworlds* and those in *cultures of risk*. On this theoretical basis, we will try to explain, how implementations of these strategies can be found in popular media products. For this, we will take a closer look at the online soccer manager game on [www.kicker.de](http://www.kicker.de) and the dating platform Parship. They are both computer based technologies that virtually mediate risks in respect to real persons and their characteristics and behaviours: soccer players on the one and potential partners on the other hand. The thesis is that both use strategies of calculating and minimizing risk according to the logic of securityworlds and of playing with risk according to the logic of cultures of risk at the same time. Further, they do their part to establish the ideas and strategies of risk and risk management in popular culture and help naturalizing the attached knowledge and practices.

### **Paper**

The scholarly perspective on the concept of security has become seemingly inevitably connected to the concept of risk during the last years. In contrast to danger, risk is something virtual that can only be applied by visualization and statistics that make it calculable and therefore manageable. Further, risk lays the responsibility for this management and the outcome of actions on the acting subject. The political scientist Herfried Münkler describes two ideal types of strategies to deal with this task:

*securityworlds* and *cultures of risk*. *Securityworlds* try to exclude danger and threat by walling off, security technologies and risk avoidance. In doing so, they also make factors of insecurity visible and produce a higher feeling of insecurity and cultures of fear. The *cultures of risk* on the other hand face dangers and threats by taking risks and having a chance in both, a playful and calculating way. The two concepts do not exclude one another but frame and presuppose each other (Münkler, 2009).

Both strategies are based on models and technologies of visualization and calculation that are mainly statistical. For storing, sorting, searching, relating and processing these numeric data, computer based databases seem to be the perfect device. They are the technical infrastructure for generating risk profiles and scenarios that are used for calculating risks and for choosing options of action. So, databases are on the one hand a tool for handling risks and on the other hand the technology that makes risk visual and the concept thinkable at first.

This connection between discourses, practices and technology is interesting because it evokes questions about the “risky” implications and inscriptions in computer databases used in everyday life in which actions and practices are being monitored permanently. Popular media like computer games or internet applications are the most influential media in the contemporary popular culture and providing “orientative knowledge” for our lives by giving “patterns of knowledge and actions”, the subject can “adapt on and accommodate” (Neitzel and Nohr 2008).

Within our presentation, we will examine in which respect the concepts of *securityworlds* and *cultures of risk* are negotiated and implemented in the popular media products [www.parship.de](http://www.parship.de) and the soccer manager game on [kicker.de](http://kicker.de) and which patterns of knowledge and action are provided in them. Both objects combine purely databased elements (personal profiles and a mathematical matrix for rating soccer players) with real world elements (real persons as potential partners and the real efforts of soccer players) in a popular medial context. In the analysis we will have a look at the different and similar strategies of risk management that try to mediate the calculability of the database and the contingency of the real world.

## References

- Münkler, H. (2009). Strategien der Sicherung: Welten der Sicherheit und Kulturen des Risikos. Theoretische Perspektiven. In: H. Münkler, M. Bohlender and S. Meurer (Eds.), *Sicherheit und Risiko. Über den Umgang mit Gefahr im 21. Jahrhundert* (pp.11-34). Bielefeld: Transcript.
- Neitzel, Britta & Nohr, Rolf F. & Wiemer, Serjoscha (2009): Benutzerführung und Technik-Enkulturation. Leitmediale Funktionen von Computerspielen. In: D. Müller, A. Ligensa and P. Gendolla (Eds.), *Leitmedien. Konzepte – Relevanz – Geschichte* (pp.231-256). Bielefeld: Transcript.

## Informational Warfare and Just War Theory

MARIAROSA TADDEO

**Abstract.** This paper focuses on Informational Warfare – the warfare characterised by the use of information and communication technologies. This is a fast growing phenomenon, which poses a number of issues ranging from the military implementation of such technologies to its political and ethical implications. The paper presents a conceptual analysis of this phenomenon with the goal of investigating its nature. Such an analysis is deemed to be necessary in order to lay the ground for future work on this topic addressing the ethical problems engendered by Informational Warfare. The analysis is developed in three parts. It first delineates the relation between Informational Warfare and the Information revolution. It then turns the attention to the effects that the diffusion of this phenomenon has on the concepts of state and war. On the basis of this analysis, it provides a definition of Informational Warfare as a *transversal* phenomenon for what concerns the environment in which it is waged, the way it is waged and the ontological and social status of the involved agents. Finally, the paper concludes taking in consideration Just War Theory and the problems arising from its application to the case of Informational Warfare.

### Extended Abstract

The analysis presented in the paper focuses on Informational Warfare (IW) – the warfare based on the use of Information and Communication Technologies (ICTs). IW has been at the centre of interest of governments, intelligence agencies, computer scientists and security experts for the past two decades (Arquilla 1999; Libicki 1996; Singer 2009). ICTs support war waging in two ways: providing new weapons to be deployed on the battlefield – like drones and semi-autonomous robots - and allowing for the so-called *information superiority*, the ability to collect, process, and disseminate information while exploiting or denying the adversary's ability to do the same.

ICTs prove to be effective and advantageous war technologies, as they are efficient and relatively cheap when compared to the general costs of war. For this reason, the use of ICTs in warfare has grown rapidly in the last decade determining some deep changes in the way war is waged, giving the rise to the latest revolution in military affairs (RMA).

This RMA concerns *in primis* military force. It also concerns strategy planners, policy-makers and ethicists, as the need to regulate this new form of warfare is much felt and the existing international regulations, like the Geneva and Heuge Conventions, provide only partial guidelines. In the same way, traditional ethical theories of war,

which should provide the ground for policies and regulations, struggle to address the ethical problems that arose with this new form of warfare (Arquilla 1999; Arquilla and Boerer 2007; DeGeorge 2003; Hauptman 1996; Powers 2004). There are three categories of problems on which both policy-makers and ethicists focus their attention, and these are the *risks*, *rights* and *responsibilities*. In the paper I will refer to these problems as to the 3R problems. Altogether, the 3R problems pose a new ethical challenge. Nevertheless, such problems will not be the focus of this paper, which will rather concentrate on the analysis of the nature of IW and the changes that it determines. The task of the proposed analysis is to lay down the conceptual foundation for the solution of the 3R problems, which will be provided in elsewhere. IW it is a wide spectrum phenomenon, which is rapidly changing the dynamics of combat as well as the role warfare in political negotiations and the dynamics of civil society. These changes are the origins of the 3R problems, the conceptual analysis of such changes and of the nature of this phenomenon is deemed to be a necessary and preliminary step to solve these problems.

The analysis is divided in three steps. First, IW is analysed within the framework of the Information revolution (Floridi 2009). Floridi's analysis of Information revolution as the fourth revolution is recalled and it is stressed that such a revolution determines a *shift toward the non-physical domain*, the domain of nonphysical objects, agents and interactions.

In the second step, it is argued that IW is one of the most compelling cases of such a shift. This analysis leads to the consideration of the effects of the dissemination of IW on the concepts of war and state. In particular, it is argued that IW redefines the concept of war as a phenomenon not necessary sanguinary and violent, and rather *transversal* for what concerns the environment in which it is waged, the way it is waged and the ontological and social status of its agents. A definition stressing the transversality of IW and its disruptive nature is then provided.

**Informational Warfare** is the use of ICTs within an offensive or defensive military strategy aiming at the disruption of the enemy's resources, and which is waged within the informational environment, by agents and targets ranging both on the physical and non-physical domains and whose level of violence may vary upon circumstances.

Finally, the third step is devoted to consider the problems arising when IW is considered within the framework of Just War Theory. This theory provides the ground for international regulations, and sets the parameters for both the ethical and the political debates. The issue is addressed whether and how the principles of Just War Theory could be applied to IW.

The analysis unveils three problems. The first one concerns the differences between the scenario assumed by Just War Theory and the one delineated by IW. Just War Theory refers to classic warfare, where governments and their leaders are the only ones who inaugurate wars by deploying armed forces, and they are the ones to be held accountable the actions of war. IW fosters a completely new way of declaring and waging war. The need is stressed for Just War Theory to take into account such changes in order to address the ethical problems arose with IW. The other two problems concern the application of two principles of Just War Theory – 'war as last resort' and 'discrimination and non-combatants immunity' – to the case of IW. In the case of the principle of 'war as last resort' the analysis indicates that the application of this principle

to the case of IW leads to an ethical impasse. The principle assumes that war is a violent and sanguinary phenomenon. It is argued that the correctness of this assumption is shaken when IW is taken into account, and that in these circumstances the application of the principle of war as last resort becomes less immediate. The impasse concerns the use of bloodless and non-physically violent modes of combat peculiar of IW, like a cyber attack, to address potentially dangerous diplomatic conflicts to prevent the occurrence of classic warfare. On one hand, such a use constitutes an act of war itself and as such Just War Theory forbids it, on the other hand it may avoid states to engage in a sanguinary war and hence is intrinsically consistent with the overall view proposed by Just War Theory of reducing bloodshed and conflicts.

A similar ethical problem is described with respect to the application of the 'principle of discrimination and non combatants immunity'. It is stressed that this principle tacitly equates non-combatants to civilians and that such an equation has been weakened by the diffusion of terrorism and guerrilla, to become even feebler with the dissemination of IW. In IW scenario, civilians may take part to a combat action from the comfort of their homes, while carrying on with their civilian life and hiding their status of informational warriors.

An ethical conundrum is described. Given the difficulty to distinguish combatants from non combatants in IW scenario, and in order to endorse the 'principle of discrimination', states might be justified to embrace high levels of surveillance over the entire population breaching individual rights, like privacy and anonymity, in order to identify the combatants and guarantee the security of the entire community.<sup>9</sup> It is argued that, on the one side, respecting the principle of discrimination may lead to violate individual rights. On the other side, waving the principle of discrimination leads to bloodshed and dissemination of indiscriminate violence over the civil population. The paper concludes pulling together the threads of the analysis and stressing the importance of developing ethical guidelines, which will provide the ground for the definition of the necessary regulation for IW and for the solution of the 3R problems.

## References

- Arquilla, J. (1999). Ethics and information warfare. Strategic appraisal: the changing role of information in warfare. Z. Khalilzad, J. White and A. Marsall. Santa Monica, USA, Rand Corporation: 379-401.
- Arquilla, J. and D. A. Borer, Eds. (2007). Information Strategy and Warfare: A Guide to Theory and Practice (Contemporary Security Studies). New York, USA, Routledge. DeGeorge, R. T. (2003). "Post-september 11: Computers, ethics and war." Ethics and Information Technology 5(44): 183-190.
- Hauptman, R. (1996). "Cyberethics and social stability." Ethics and Behavior 6(2): 161-163.
- Floridi, L. (2009). "The information Society and Its Philosophy." The Information Society 25(3): 153-158.
- Libicki, M. (1996). What is Information Warfare? Washington, D.C, USA, National Defense University Press.

---

<sup>9</sup> This problem is part of the 3R problems described in section one.

- Powers, T. M. (2004). "Real Wrongs in Virtual Communities." *Ethics and Information Technology* 5(4): 191-198.
- Singer, P. W. (2009). "Robots at War: The New Battlefield." *Wilson Quarterly* 33(1): 30-48.

## **TECHNO-SECURITY, RISK AND THE MILITARIZATION OF EVERYDAY LIFE**

JUTTA WEBER

*University Paderborn*

*Warburger Straße 100, 33098 Paderborn*

**Abstract.** Recently, we experience a rapid and ongoing transfer of security technologies such as body scanners, drones, or biometrics from the military realm in everyday life. And though there is a lively debate on the growing militarization of public space, political culture and everyday life (Giroux 2004, Graham 2005, Crandall/Armitage 2005, Kohn 2009) there is surprisingly little discussion on the huge amount of military-civilian transfer of new and emerging security technologies. Only very few authors address the possible militarization of society through the procurement, adaptation and proliferation of military technologies in civilian life (Agre 2001). A few scholars such as Dandeker (1990, 2006), Wood et al. (2006), or Balzacq et al. (2010) pointed out that security technologies and practices are deeply impregnated by their military offspring. Surveillance studies scholars – leaning on Anthony Giddens (1985) – at least partly acknowledge the growing entanglement of the military and bureaucracy in post/modern societies (Bogard 1996, Dandeker 1990, Nellis 2009, Wood et al. 2003). Approaches in STS (Akrich 1992; Woolgar 1991) and philosophy of technology (Winner 1986, Verbeek 2006, Flanagan, Howe and Nissenbaum 2006) showed how technology transports values, world views and norms. Therefore I will ask in my paper what norms, values, frames of thought are transported into everyday life with the military-civil transfer of security technologies – for example when uninhabited aerial vehicles become part of everyday experiences for example through the growing presence of UAVs during global sport and cultural events, by demonstrations or during law enforcement as well as through ‘augmented reality video games’.

### **2. Daily Drones. Techno-Security & the Militarization of Everyday Life**

Originally hopes of a large-scale military-civilian conversion arose after the end of the cold war. But these hopes were disappointed already in the early 1990s when force has become again a frequent tool of foreign policy concentrating on so-called rogue and failed states that followed a growing number of military responses from peace-keeping operations up to massive invasions (Rappert et al. 2008). In philosophy of technology as well as science and technology studies (STS) we got some studies on the crossover of global communication and military surveillance systems (i.a. de Landa 1991, Edwards 1996) as well as the fusion of military, industry and media (Der Derian 2001,

Lenoir/Lowood 2002). The shift of the business of major arms manufacturers towards mainstream security and surveillance products in the post-cold-war era is addressed (i.a. Wood et al 2006, Eick 2010, Graham 2010).

Nowadays new products are developed and partially already deployed. Think of non-lethal weapons, i.a. electroshock and heat-ray weapons, as well as monitoring systems linked to killing or paralyzing systems. These weapons for warfare respectively crowd control are situated between the military and civilian realm. In a brochure on new security projects in the 7th framework programme for research, the Directorate General Enterprise and Industry of the EU commission states: “Moreover, the relationship between defence technologies on the one hand, and security technologies on the other, is particularly noticeable in the field of R&D, with technologies that show potential developments in both areas (Dual Use). At both research and industrial development levels, *synergies are possible and desirable.*” (European Commission. Enterprise and Industry 2009, my emphasis). Contemporary surveillance studies also point towards the close relation between the military and the managerial: “Cross-fertilization between the military and the managerial is clearly central to problems and developments in the study and practice of surveillance...” (Wood et al. 2003, 146). But there are very few studies on the relation of the sociotechnical, political, and the military with regard to military-related security technologies and their impact on everyday life.

## 2.1. TECHNO-SECURITY, RISK AND UNPREDICTABILITY

So what to think of the manifest development expansion of military technologies in civilian life in general and of UAVs specifically? For a long time we know about the conversion and adaptation of military technology in everyday life – think only of recent examples of the military offspring of technologies such as the internet, RFID, satellite technology or GPS (Global Positioning System). Approaches in STS (Akrich 1992; Woolgar 1991) and philosophy of technology (Winner 1986, Verbeek 2005, Flanagan, Howe and Nissenbaum 2006) showed how technology transports values, world views and norms. Madeleine Akrich made visible that every technology contains scripts while Steve Woolgar (1991) pointed to the fact that technology is „configuring the user“ and the context of the use. Therefore it is important to ask which frames of thought, world views, perspectives, preferences and motives are inscribed into military-related security technologies and translated into everyday life. Kaplan (2006) has shown how GPS did not only link demography, geography, remote sensing, geopolitics and identity politics but how GPS became an icon of “personal empowerment and self-knowledge linked to speed and precision” (Kaplan 2006: 697) for US Americans. At the same time the „militarized consumer“ who wants to improve his „lifestyle“ provides the personal data thereby enabling new systems of surveillance (embedded in mobiles, GPS systems in cars, etc.): “...tracked, the user becomes a target within the operational interfaces of the marketing worlds, into whole technologies state surveillance is outsourced.” (Crandall 2006, np)

Relevant epistemological shifts and the emergence of new norms, worldviews and values that accompany the massive contemporary military-civilian transfer is the epistemological reframing of today’s concept of security. Homeland as well as international security is not primarily occupied with the defense against specific threats and prosecuting crimes (Albrecht 2009) but with the (precautionary) management of risk and preventive and pre-emptive securitization of security (Aradau et al. 2008, Ammicht-Quinn/Rampp 2009, Zedner 2007). While traditionally threat was related to actions and intentions of conflicting parties which can be – in principle resolved, the concept of „risk“ embrace the idea of general, permanent and systemic contingencies such as pandemics, global warming, rogue states, terrorism, organized crime, poverty, illegal

immigration or the proliferation of weapons of mass destruction (European Commission. Enterprise and Industry 2009). The concept of risk is closely entangled with *unpredictability* and *insecurity* – especially with regard to the identification of the enemy or the assessment of hazardous situations. *The politics of risk operates with risk profiling on the basis of statistics and probabilities, with models and speculations which do not target at eliminating but managing risk*: „In short, whereas the concept of threat brings us in to the domain of the production, management and destruction of dangers, the concept of risk mobilizes and focuses on different practices that arise from the *construction, interpretation and management of contingency*“ (Aradau et al. 2008, 148; my emphasis) This new approach is highly technological-oriented. The shift towards a preventive security policy and a techno-centred concept of security corresponds to the increasing networking of surveillance measures. The reconfiguration of surveillance as assemblage (Haggerty/Ericson 2000) is a general tendency. Nevertheless, the concept and practice of digital network-centred surveillance technologies (Graham/Wood 2003) shows strong affinities to that of network-centric warfare. The latter – also called „Revolution in Military Affairs“ – is based on strong, ubiquitous ICT-based networks and mobilities that control and monitor area-wide and over huge distances 24 hours a day to reach a “globespanning dominance based on a nearmonopoly of space and air power (Graham 2005, 175; see also Dillon 2002, Dandeker 2006). In this scenario, especially autonomous UAVs with artificial intelligence and learning capability are regarded as an important component of new techno-warfare (Weber 2009, 2010). Together with inhabited systems integrated in a complex network of air, water and ground agents, new techniques of warfare are developed “... toward a vision of a strategic and tactical battlespace filled with networked manned and unmanned air, ground, and maritime systems ... that free warfighters from the dull, dirty, and dangerous missions ... and enable entirely new design concepts unlimited by the endurance and performance of human crews. The use of UAVs in Afghanistan and Iraq is the first step in demonstrating the transformational potential of such an approach.” (Department of Defense 2007, 34) This aspired high-tech transformation of armed forces is supposed to make them invincible, to develop strategies of digital deterrence more powerful than nuclear deterrence ever was. The *utopia of a ubiquitous, networked system of surveillance and control* seems to be mirrored by a preventive and techno-centred idea of security in everyday life – for example when drones are deployed for law enforcement by the British Police or for border control by the European agency Frontex.

Recently, the Guardian’s Freedom of Information request revealed the very broad scope of potential UAV applications by the British police: “Working with various policing organisations as well as the Serious and Organised Crime Agency, the Maritime and Fisheries Agency, HM Revenue and Customs and the UK Border Agency, BAE [systems; the British defence company] and Kent police have drawn up wider lists of potential uses. One document lists ‘[detecting] theft from cash machines, preventing theft of tractors and monitoring antisocial driving’ as future tasks for police drones, while another states the aircraft could be used for combat ‘fly-posting, fly-tipping, abandoned vehicles, abnormal loads, waste management’ (...) There are two models of BAE drone under consideration, neither of which has been licensed to fly in non-segregated airspace by the CAA. The Herti (High Endurance Rapid Technology Insertion) is a five-metre long aircraft that the Ministry of Defence deployed in Afghanistan for tests in 2007 and 2009”. (Lewis 2010).

According to these plans, the use of UAVs would be part of a larger network-centric project through which information from a variety of sources (UAVs, smart CCTV, data detention, analysis of money transfer, etc.) are networked and evaluated. This course of action seems not to aim primarily at prosecuting specific crimes and

follow concrete suspicions but *to search monitor a nation's population systematically and thoroughly on an everyday basis*. We need to investigate whether this civilian approach resembles what is called C4ISR – Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance in the military. C4ISR stands for the networking of all available surveillance and control systems to achieve a global overview in the war theatre. So maybe we witness the idea of a global overview in the (civilian) world theatre.

Part of these epistemological and normative reframing might also be found in recent consumer applications of UAVs. Since last year the first little UAVs respectively quadricopters are available for „augmented reality video games“ (<http://ardrone.parrot.com/parrot-ar-drone/de/>) in which one can launch missiles and fight against other drones. The quadricopters can be controlled by an iPhone, iPod Touch or iPad. There are two cameras embedded into the drone, one on the front and one underneath, to enable a direct sight via video remote control on the basis of a Wi-Fi connection. Another application is provided by a German company which rents drones for private use ([www.rent-a-drone.de](http://www.rent-a-drone.de)) to enable real time pictures and videos from above.

The private consumer applications of UAV might (still) not be as wide ranging as GPS but in a way one could argue that they might open the door in more intense participatory surveillance and observation practices (Ball 2005, Koskela 2009). Daily consumer drones might contribute to train users to watch the world from a top-down or „God's eye view“ that participates in the C4ISR longing for a global overview in the war / world theatre.

The tightening networks of surveillance technologies – increasingly expanded by drones for border control, policing demonstrators, crowd and event control, are part of a growing belief in “smart”, specific, side-effects-free, information-driven utopia of governance” (Valverde and Mopas, 2004: 239). Network centric warfare with its idea of C4ISR relies on this utopia as it might be the case with recent police applications of drones and new gamer applications such as the iPhone controlled ar-drone. It is necessary to follow up closely the growing transfer of military technologies in civil applications, game practices and other everyday life to see whether and how recent ideas of techno-security and „full spectrum dominance“ become dominant in 21st century's societies of control.

## References

- Ammicht-Quinn, R. & Rampp, B. (2009). The Ethical Dimension of Terahertz and Millimeter-Wave Imaging Technologies – Security, Privacy and Acceptability: Optics and Photonics. In: C.S. Halvorson et al. (Eds), *Global Homeland Security V and Biometric Technology for Human Identification VI* (pp. 1-11). Proc. of SPIE Vol. 7306, 730613.
- Agre, P.E. (2001). *Imaging the Next War. Infrastructural Warfare and the Conditions of Democracy*. Retrieved from <http://polaris.gseis.ucla.edu/pagre/war.html> [accessed 17 November 2010].
- Akrich, M. (1992). The de-scription of technological objects. In: W.E. Bijker & J. Law (Eds.), *Shaping technology/building society* (pp. 205-224). Cambridge: MIT.
- Ammicht-Quinn, R. & Rampp, B. (2009). The Ethical Dimension of Terahertz and Millimeter-Wave Imaging Technologies – Security, Privacy and Acceptability: Optics and Photonics. In: C.S. Halvorson et al. (Eds), *Global Homeland Security V and Biometric Technology for Human Identification VI* (pp. 1-11). Proc. of SPIE Vol. 7306, 730613.

- Ball, Kirstie Ball (2005). Organization, Surveillance and the Body: Towards a Politics of Resistance. In: *Organization*. Volume 12(1): 89–108
- Balzacq, T. et al. (2010). Security Practices. In: R. Denemark (Ed.), *International Studies Encyclopedia Online*. Retrieved from <http://didierbigo.com/documents/SecurityPractices2010.pdf> [accessed 4 November 2010].
- Bogard, W. (1996). *The Simulation of Surveillance: Hypercontrol in Telematic Societies*. Cambridge: Cambridge University Press.
- Capurro, R., Tamburrini, G. & Weber, J. (Eds.) (2008). *Techno-Ethical Case-Studies in Robotics, Bionics, and Related AI Agent Technologies. Deliverable 5 of the EU-Project ETHICBOTS. Emerging Technoethics of Human Interaction with Communication, Bionic and Robotic Systems (SAS 6 - 017759)*. Retrieved from <http://ethicbots.na.infn.it/restricted/doc/D5.pdf> [accessed 17 November 2010].
- Crandall, J. & Armitage, J. (2005). Envisioning the Homefront: Militarization, Tracking and Security Culture. *Journal of Visual culture*. 4 (1), 17-38.
- Crandall, Jordan (2006). *Operational Media*. Retrieved from <http://www.ctheory.net/printer.aspx?id=441> [accessed 2nd January 2011].
- Dankeker, C. (1990). *Surveillance, Power and Modernity: Bureaucracy and Discipline from 1700 to the Present Day*. New York: St. Martin.
- Dandeker, C. (2006). Surveillance and Military Transformation: Organizational Trends in Twenty-first-Century Armed Services. In: K.D. Haggerty & R.V. Ericson (Eds.), *The new politics of Surveillance and Visibility* (pp. 225-249). Toronto, Buffalo and London: University of Toronto Press.
- De Landa, M. (1991). *War in the Age of Intelligent Machines*. New York: Zone Books.
- Department of Defense (2007). *Unmanned Systems Roadmap 2007-2032*. Retrieved from <http://www.acq.osd.mil/usd/Unmanned%20Systems%20Roadmap.2007-2032.pdf> [accessed 12 June 2008].
- Der Derian, J. (2001). Virtuous war: mapping the military-industrial-media entertainment network, *Westview Press*, Boulder, CO.
- Edwards, P.N. (1996). *The Closed World: Computers and the Politics of Discourse in Cold War America*. Cambridge, MA: MIT Press.
- Eick, V. (2010). *The Droning of the Drones. The increasingly advanced technology of surveillance and control*. Retrieved from <http://www.statewatch.org/analyses/no-106-the-droning-of-drones.pdf> [accessed 12 November 2010].
- European Commission. Enterprise and Industry (2009). *Security Research. Towards a more secure society and increased industrial competitiveness. Security Research Projects under the 7th Framework Programme for Research*. May 2009. Retrieved from [ftp://ftp.cordis.europa.eu/pub/fp7/security/docs/towards-a-more-secure\\_en.pdf](ftp://ftp.cordis.europa.eu/pub/fp7/security/docs/towards-a-more-secure_en.pdf) [accessed 17 November 2010].
- Flanagan, M., Howe, D.C. & Nissenbaum, H. (2008). Embodying Values in Technology. In: J. van den Hoven & J. Weckert (Eds.), *Information Technology and Moral Philosophy* (pp. 322-353). Cambridge: Cambridge University Press.
- Giddens, A. (1985). *The Nation-State and Violence. A Contemporary Critique of Historical Materialism*, Vol. II. Berkeley: University of California Press.
- Giroux, H.A. (2004). War on Terror. The Militarising of Public Space and Culture in the United States. *Third Text*. Vol. 18, Issue 4, 211-221.

- Graham, S. (2005). Surveillance, urbanization and the US „Revolution in Military Affairs“. In: D. Lyon (Ed.), *Theorizing Surveillance. The panopticon and beyond* (pp. 247-270). Devon, UK: Willian.
- Graham, S. & Wood, D. (2003). Digitizing Surveillance: Categorization, Space, Inequality. *Critical Social Policy*. Vol. 23, No. 2, 227-248.
- Graham, S. (2010). *From Helmand to Merseyside: Unmanned drones and the militarization of UK policing*. Retrieved from <http://www.opendemocracy.net/ourkingdom/stevegraham/from-helmand-to-merseyside-military-style-drones-enter-uk-domestic-policing>, [accessed 17 November 2010].
- Haggerty, K. & Ericson, R. (2000). The surveillance assemblage. *British Journal of Sociology*. Vol. 51, No. 4, 605-622.
- Kaplan, Caren (2006). Precision Targets: GPS and the Militarization of U.S. Consumer identity. *American Quarterly* 58.3 ,693-713.
- Koskela, Hille (2009). Hijacking surveillance? The new moral landscapes of amateur photographing. In: Katja Franko Aas, Helene Oppen Gundhus, Heidi Mork Lomell (Eds.) *Technologies of Insecurity: The Surveillance of Everyday Life*. (pp.147-168). Oxon / New York: Routledge-Cavendish.
- Kohn, R.H. (2009). The Danger of Militarization in an Endless „War“ on Terrorism. *The Journal of Military History*, Vol. 73, No. 1, 177-208.
- Lenoir, T. & Lowood, H. (2002). *Theaters of War: The Military-Entertainment Complex*. Retrieved from [http://www.stanford.edu/class/sts145/Library/Lenoir-Lowood\\_TheatersOfWar.pdf](http://www.stanford.edu/class/sts145/Library/Lenoir-Lowood_TheatersOfWar.pdf) [accessed 17 November 2010].
- Lewis, P. (2010). CCTV in the sky: police plan to use military-style spy drones. *The Guardian* (London), 23.1.2010. Retrieved from [www.guardian.co.uk/uk/2010/jan/23/cctvsky-police-plan-drones](http://www.guardian.co.uk/uk/2010/jan/23/cctvsky-police-plan-drones) [accessed 12 November 2010].
- Nellis, M. (2009). 24/7/365: mobility, locatability, and the satellite tracking of offenders. In: Katja Franko Aas, Helene Oppen Gundhus, Heidi Mork Lomell (Eds.) *Technologies of Insecurity: The Surveillance of Everyday Life*. (pp.103-124). Oxon / New York: Routledge-Cavendish.
- Rappert, B., Balmer, B. & Stone, J. (2008). Science, Technology and the Military. Priorities, Preoccupations and Possibilities. In *The Handbook of Science and Technology Studies*. London: MIT Press, 719-740.
- Verbeek, P.-P. (2006). Materializing Morality. Design Ethics and Technological Mediation. *Science, Technology & Human Values*. Vol. 31, No. 3, 361-380.
- Weber, J. (2009). Unmanned Combat Aerial Vehicles, Dual Use and the Future of War. In: R. Capurro, M. Nagenborg & G. Tamburinni (Eds.), *Ethics and Robotics*, (pp.83-103). Amsterdam/Heidelberg: IOS Press: Deutscher Akademieverlag.
- Weber, J. (2010). Armchair Warfare „on Terrorism“. On Robots, Targeted Assassinations and Strategic Violations of International Law. In: Jordi Vallverdú (Ed.): *Thinking Machines and the Philosophy of Computer Science: Concepts and Principles* (pp.206-222). IGI Global.
- Winner, L. (1986). *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. Chicago: University of Chicago Press.
- Woolgar, S. (1991). Configuring the User: The Case of the Usability Trails. In: John Law (Ed.), *A Sociology of the Monsters. Essays on Power, Technology and Domination*. (pp.59-99). Verlag: London and other, 59-99.

# **Track V: Information Ethics, Robot Ethics**

## IS THERE A HUMAN RIGHT NOT TO BE KILLED BY A MACHINE?

PETER M. ASARO  
*The New School University*  
*asarop@newschool.edu*

### 1. Extended Abstract

This presentation reviews the standard frameworks for considering the human right not to be killed, and its forfeit by combatants in a war. It then considers as a special case the right not to be killed by a machine. Insofar as one has a right not to be killed by any means, then one also has a right not to be killed by a machine, such as a lethal robotic system. It is further argued that in those cases in which an individual may have already forfeited their right not to be killed, such as when acting as a combatant in a war, this does not necessarily subject one to being killed by a machine. Despite a common view that combatants in war may be liable to be killed by any means, “killing by machine” fails to meet the requirements for ethically justifiable killing. The defense of this assertion will rest on a technical definition of “killing by machine,” and further clarification of justified killing in war. In short, the argument is that “killing by machine” fails to consider the rights of an individual in the morally required manner. This is because “killing by machine” requires a “decision to kill” to be made by a moral agent, and an automated decision cannot involve the necessary moral deliberation required to justify violating the human right not to be killed. As such, automated decisions to kill are not morally justifiable.

The argument begins by examining the right to self-defense which forms the rights-based interpretation of Just War Theory. In particular, I examine the “Castle Laws”, aka “Make My Day Laws,” which permit individuals to use force against home-intruders without criminal or civil liability in many U.S. states. I examine the conditions under which individuals in such circumstances are permitted to use lethal force, and when such force becomes “willful and wonton misconduct.”

Informed by this analysis, I examine the legality of a home-defense robot, and the legal permissibility of its use of force against home-intruders. In general, the “Castle Laws” do not allow homeowners to booby-trap their homes, and a robotic home-defense system can be viewed as a sophisticated booby-trap. I consider the various objections that might be made to the standard rejection of booby-trap. According to such objections, a robot with sophisticated cognitive and perceptual capabilities might be argued to avoid manifesting a form of “reckless endangerment.”

I then analogize from the case of home-defense in civil and criminal law, to the case of self-defense in war, and the Laws of Armed Conflict and Just War Theory. While warfare has much looser standards of what constitutes a “threat,” and the proximity of threats, the use of systems capable of automated lethal decision-making is largely analogous to the domestic use of booby traps.

I conclude that implicit in both domestic law and international laws of armed conflict is requirement for moral deliberation which undermines the moral and legal legitimacy of automated lethal decision making. This has serious implications for the use of autonomous lethal robotics in police and military applications. One implication is that only artificial moral agents, capable of exercising moral autonomy, could be morally and legal justified in violating the rights of a human.

## DO WE NEED AN UNIVERSAL INFORMATION ETHICS?

THOMAS CHRISTOPHER DASCH  
*University of Paderborn*  
*Germany*

**Abstract.** This article deals with information ethics. This raises the essential question: What is information? But I want to focus on the ethical category. Herefore, three areas of potential actions arise. Instead of informations I want to talk more generally of data. This makes it possible to distinguish between: (1) The pure receive of data, (2) The pure provision of data, (3) The simultaneous receive and provision of data, (4) A further possible action is to supply a platform for data. This is strictly speaking the topic three, but it will be discussed as an separate topic. Here is exemplified the ethical problems for the individual cases may occur. Subsequently, a connection between the problems of the legislation of the Internet and the lack of a universal ethical base is made in the information ethics.

This article deals with information ethics. This raises the essential question: What is information?

The question of "What is Information?" (Floridi, 2004, p.560) is according to Floridi the elementary problem of the philosophy of information. Among the advocates of well known approaches to the concept of information are Shannon and Weaver, Bar-Hillel and Carnap, Wiener, Janich, etc. (Capurro, 2000). Here Capurro's trilemma (Fleissner, Hofkirchner, 1995) applies: (1) Either the concept of information is always the same no matter what the set of input data is like, (2) or the information is only of similar kind, or (3) it is completely independent. At this point it is to be clarified on which concept of information based the information ethics.

But I want to break another ground. I want to focus on the ethical category. In this context information ethics is the part of ethics that deals with the internet. The concept of information is to be ignored here. "Morale is focussed on judgments, that assess a human action positively or negatively, approve or disapprove it." (Birnbacher, 2007, p.12). Therefore, three areas of potential actions arise. Instead of informations I want to talk more generally of data. This makes it possible to distinguish between:

1. The pure receive of data
2. The pure provision of data
3. The simultaneous receive and provision of data
4. A further possible action is to supply a platform for data. This is strictly speaking the topic three, but it will be discussed as an separate topic.

One example for the first topic is the reading of news pages or blogs. In this context, the information content the receiver consumes is moral relevant. A possible

moral misconduct in this field is the download of music without owning the respective rights. In case of the internet, the information recipient may not be able to reconstruct the origin of the information. Additionally, the information can be deleted from the respective homepage at any time. In contrast, the information content of a newspaper can not be changed once the paper is printed.

The second topic includes e.g. owners of news pages. In this connection, the precise content of the online data is moral relevant. In the case of news pages it is expected that the news have been extensively investigated. One example for the misuse of this function is a scenario in which a person spreads videos showing another person in an unfavourable context. In the case of the internet, tracking down the owner of the page is far more difficult as tracking down a normal information transmitter. The latter differs from the internet in concerns of judicial matters, more about that later in the text. A feature of the internet is that a large group of people can be addressed without the need of a major news infrastructure. Interest groups can be formed rapidly and easily in this way as seen recently when a open letter was handed to Chancellor Merkel concerning the plagiarism affair of Germanys minister of defence, Karl Theodor zu Guttenberg. In this way, the initiators of the letter were able to support the ministers retirement.

Amongst others, topic three includes chats, forums and online games. In this case, moral relevance is similar to moral relevance in non virtual communication. A possible moral misconduct would e.g. be the insult to a person in a chat room. Characteristic for this kind of online communication is that the counterpart can not be visualized (as long as webcams are not used). Therefore, it remains unknown what emotions the counterpart expresses.

“Emotions are responses of an organism centered on experiences. They represent the relevance of an artefact of perception for the fulfilment of needs (e.g. according to the criteria “beneficial” or “impedimental”). Additionally, they activate or constrain various cognitive and motivational systems in terms of a optimal satisfaction of needs.”(Kuhl, 2010, p.543) This can lead to a incorrect estimation of the counterparts emotions. However, the chatter can manipulate emotions by the use of e.g. smilies, that do not represent his actual emotions. In case of the internet, the identity of the person one is chatting with can not be verified. The counterpart is not necessarily regarded as a person, but in a distinct role. This can be the case in online games as required participant, in forums as disposer of information and so on.

The fourth topic includes for example provider or platforms like Facebook or search engines like Google and file sharing services. At this point it is ethical relevant whether the suppliers can assure a ethical correct mode for the users. An Examples for an ethical dubious action in this topic are to run a file sharing service for music without having the copy rights. A point at issue is Wikileaks, too. It is questionable, wheter it is ethically to publish diplomatic cables.

Despite all this potentially ethical critical topics one can point out that beyond this controversial concepts and opinions exist. This depects for example in the five cultural deminsions of Hofstede: Power Distance Index(PDI), Individualism(IDV), Masculinity(MAS), Uncertainty Avoidance Index (UAI), Long-Term Orientation (LTO) (Lüsebrink, 2005, p. 20-25). On the one hand this is due to different opinions about this in the respective culture area. On the other hand, different cultures show different behaviour on the internet, that can be reduced to the fact that violation on the internet against ethical basic principles remains largely unpunished. The internet is no area

immune from law, but it is so that people on the Internet are global and there depending on each of the legislation and the enforcement of the laws of their own country. “The almost traceless variability of content presents new challenges to the reliability of documents and the evidence. The indifference of original and copy has a new copyright quality. The anonymity of the web makes it difficult to identify reliable contractors. The speed of interactive communication such as short natural cooling-in contracts considerably, giving the consumer a new dimension. “(Haug, 2010, p.9)

It would require a common ethical base in information ethics.

## References

- Birnbacher, D. (2007). *Analytische Einführung in die Ethik*, 12. Berlin: Walter de Gruyter
- Capurro, R. (2000). *Einführung in den Informationsbegriff*. Available at <http://www.capurro.de/infovorl-kap3.htm> [15.02.2011].
- Fleissner, P., Hofkirchner, W. (1995). Informatio revisited. Wider den dinglichen Informationsbegriff. *Informatik Forum*, 9(3), 126-131.
- Floridi, L. (2004). Open Problems in the Philosophy of Information. *Metaphilosophy*, 35(4), 554-582.
- Haug, V. (2010). *Internetrecht: Erläuterungen mit Urteilsauszügen*, 9. Stuttgart: Kohlhammer.
- Kuhl, J. (2010). *Lehrbuch der Persönlichkeitspsychologie: Motivation, Emotion und Selbststeuerung*, 543. Göttingen: Hogrefe.
- Lüsebrink, H. (2005). *Interkulturelle Kommunikation*, 20-25. Stuttgart: Metzler.

## A PSEUDOPERIPATETIC APPLICATION SECURITY HANDBOOK FOR VIRTUOUS SOFTWARE”

KEITH DOUGLAS

*Statistics Canada*<sup>10</sup>

In the past 10 or 15 years an increased awareness of application security<sup>11</sup> (AS) in computing and information systems has resulted in many volumes of material (e.g., Cross 2006, Burnett 2004, Seacord 2005, Clarke 2009). Security conscious developers, testers, and organizations wishing to adopt “best practices” have a lot of work to distill these many volumes of advice and principles into easily implementable and understandable approaches. Following the off-hand suggestion from a colleague (Perkins 2010), I have taken her phrase “virtuous software” as a starting point. In this paper, I comb through the *Nicomachean Ethics* (Aristotle 1984) to find appropriate guidance for virtue in AS. It thus is addressed both to computing professionals wanting to understand why AS makes the ethical consequences of their work more salient (or, more debatably<sup>12</sup>, makes them exist) and also to philosophers who may not be aware of the ethical challenges raised by recognition of AS in computing. It is also intended as a brief introduction as to why AS considerations matter as one (not independent of the others) aspect of the “architecture”, design, development, and support of software.

---

<sup>10</sup> Author affiliation for identification purposes only.

<sup>11</sup> AS is to be distinguished in discussions of computing security from infrastructure security, dealing with antimalware solutions, public key utilities, routing rules in networks, etc. 70% of current exploits and vulnerabilities are in application areas (Sykora 2010) and subsequently AS merits philosophical and computational attention. It is often discussed in the context of “application hardening”. This term is in the author’s view unhelpful, since it suggests, wrongly, that a correct approach to would be to implement an application and then “fix it up” to meet the hardening requirements. The expert consensus seems to be that AS ought to be part of the entire software development life cycle, and have a role to play at almost every phase. See, e.g., Seacord 2005. The case of what to do about existing systems is more complicated; I do not address it as much in the present work, though much of what we can tease out of (or be reminded by) Aristotle applies regardless

<sup>12</sup> Conversations with colleagues on the part of the author suggest (he has not done formal investigations) that many computing professionals do not think their profession and activities raise any additional or different ethical considerations beyond those common to all humans in general or all relevant employees of a given organization. (For example, fellow computing colleagues of the author are certainly aware of their obligations under the relevant public service legislation, but do not see (for example) buffer overruns and race conditions as leading to possible ethically relevant situation. At best they are regarded as “another sort of bug”.) Further work (beyond the present one) to institute AS “consciousness” in developers will have to deal with this situation.

Philosophical topics I will briefly address in the above fashion are: the nature of technology, the nature of virtue, how virtue may be obtained, who is virtuous, what results from being virtuous and examples of what specific virtues are. All of these can be topics for complete presentations in their own right: I bring them up to simply show the rich areas of further possible investigation, and, in some cases, the pitfalls of using a “virtues framework” when it comes to software.

The philosophical topics in turn relate (here I do not indicate how, merely enumerate what will be discussed) to the following more directly computing considerations: the nature of computing professions, systems specifications, how one should learn about AS, characteristics of good software systems, how to adjudicate between AS and other design goals, how to get developers to be AS-aware and others.

Finally, I include this paper as a way of linking three phases of the so-called computational turn: the past: traditional philosophy (e.g., Aristotle); the present, the CAP conferences where computing and philosophy, traditional and otherwise is largely (but not exclusively) academic (yet fruitfully interacting), and the future, where work from CAP is also of importance to those outside. I do not suggest that these three phases are the only way to understand the historical development of the computing and philosophy movement, nor do I suggest that there has not been anything useful in the past to those outside of academia, merely that there is ample room within the topic of AS to address such considerations.

## References

- Aristotle. 1984. “Nicomachean Ethics”. In *The Complete Works of Aristotle*, vol. 2 (ed. Jonathan Barnes). Princeton: Princeton University Press.
- Burnett, Mark. 2004. *Hacking the Code: ASP.NET Web Application Security*. Burlington: Syngress.
- Clarke, Justin. 2009. *SQL Injection Attacks and Defense*. Burlington: Syngress.
- Cross, Michael. 2006. *Developer’s Guide to Web Application Security*. Burlington: Syngress.
- Perkins, Evelyn. 2010. Unpublished comment, meeting of the Secure Coding Practices Working Group, Statistics Canada.
- Seacord, Robert. 2005. *Secure Coding in C and C++*. New York: Addison-Wesley Professional.
- Sykora, Boleslav. 2010. Lecture Material, Learning Tree International Course 940.

## THE CENTRAL PROBLEM OF ROBOETHICS: FROM DEFINITION TOWARDS SOLUTION

DANIEL DEVATMAN HROMADA

*Université Paris 8 / École Pratique des Hautes Études / Lutin Userlab*

*hromi@kyberia.sk*

**Abstract.** The central problem of roboethics is defined as such: on one hand, robotics aims to construct entities which will transcend the faculties of human beings, on the other hand, some unethical acts should be made impossible to execute for such artificial beings. It can be illustrated on the case of full-fledged AI which is able to reprogram itself, or program other AIs but only in a way that the result shall not lead to the infraction of moral imperatives held by its human conceptors. Thus a programmer of such a system is posed between Skylle of his “aim to conceive an artificial entity able to do almost everything, and more efficiently than a human being” and a Charybde of “the principle of precaution commanding him to constraint the behaviour of such an entity in a way that it would never be able to execute certain acts, like that of a murder, for example”. Therefore the central problem can be also perceived as a form of solution to the problem of trade-off between the amount of “autonomy” of an artificial agent and the extent to which the “embedded ethical constraints” determine the agent’s behaviour. Believing that such a trade-off could be found, our proposal is conceived as a four-folded hybrid “separation of powers” model within which the final output to the solution of ethical dilemma is considered to be the result of mutual interaction of four independent components: 1) “Moral core” containing hard-wired rules analogous to Asimov laws of robotics 2) “Meta-moral Imperative” logically equivalent to Kant’s categoric imperative 3) “Ethico-legal codex” containing an extensible set of normative procedures representing the laws, moral norms and customs present in or induced from agent’s surroundings 4) “Mytho-historical knowledge base” grounding the agent’s representation of « possible states of the world » in the corpora of human generated myths & stories Finally, we will argue that our proposal of two induced & two embedded modules vaguely corresponds to the human morality faculty since it takes into account both its “innate” as well as “acquired” components.

## 1. Definition of the Central Problem

It may be stated that the ultimate goal of Artificial Intelligence is, for its most radical proponents like (Kurzweil, 2000; Vinge, 1993) □ □, the conception of an artificial system able to transcend all faculties nowadays attributed to human being. In accord with Turing's pioneer proposal (A. M. Turing, 2008) □ □, such proponents do not ask metaphysical questions like "Can machine have consciousness?" nor do they bother much with arguments like that of "chinese room" □ (Searle, 1982) □. More concretely: such radical engineers do not ask questions "whether faculty X can be simulated by algorithmic means", they simply take the affirmative answer as granted, and, in consequence, pose a question "how can I simulate the faculty X by algorithmic means?"

Let's define "the faculty of moral reasoning" as  $X_1$ . While being aware that nothing really proves that such a definition does NOT result in a fallacy, we nonetheless do not ask whether it makes sense or not to speak about "machine endowed with morality". The fact that machines will be able, sometimes in the future, able to fully simulate the moral reasoning is taken as granted within the scope of our Gedankenexperiment and the question which is posed hereby is therefore "how could it be done?"

Now let's define "the ability to modify itself" as  $X_2$  and "the ability to reproduce" as  $X_3$ . Since  $X_1$ ,  $X_2$  and  $X_3$  are all faculties commonly attributed to human being, it can be stated that an artificial system endowed with such faculties would seem more "human" than the one which contains only some of them, and is therefore closer to ultimate goal of radical AI as was already defined.

The problem arises when one realises that  $X_1$  is not necessarily mutually consistent with  $X_2$  or  $X_3$ . Myths as well as history itself demonstrate far too often to pass that the modification or a reproduction of a moral being does not necessarily yield a moral result. It is verily this "lesson from history" that obliges us to postulate the central problem of roboethics :

*How could (the most radical of) roboengineers possibly conceive a machine which is, in the fullest possible extent, able to adapt itself to any situation whatsoever and yet "unable" to rewrite the set of moral imperatives with which it was endowed ?*

We exclude completely the possibility of not endowing a machine with any moral reasoning at all. Not only would a deployment of such a self-copying, self-modifying autonomous agent be contrary to precautionary principle (Andorno, 2004) □ □, but the very intention of "creating a machine analogous in all its functions to human being" would miss its target since it is commonly accepted fact that the faculty  $X_1$ , i.e. morality is one among such anthropological universalia (Mikhail, 2007) □ □.

What's more, according to Kant - who analysed the faculty of morality and its relations to other forms of reasoning in such an extent that his discoveries simply have to be taken into consideration by anyone aiming to embed morality into machines -  $X_1$  is not only "one faculty amongst many", but it occupies the central place among all the faculties with which a man was endowed. For Kant, man is conceived, as a "moral being" (Kant, 1785) □ □.

Being moral means simply to be able to find a "good" solution to any situation of moral dilemma whatsoever. Therefore, any advanced implementation of morality into an artificial agent should not ignore the semantic intricacies of the concept of "good" nor its strong cultural and contextual dependence (i.e. what is good in one context is not necessarily good in other).

## 2. Possible solution to Central Problem

The Hebbian network of semantic relations around the term “good” consists the outermost layer of our 4-component model of a so-called “moral machine” (MM).

Initially, this graph-like structure of semantic relations could be possibly built by means of extraction of “morals of the stories” from huge hypertext corpora representing the myths, fairy tales and descriptions of factual historical situations (inputs) and their consequences (outputs).

Whether association of such inputs & outputs by means of already existing machine learning procedures (ANN, SVM, boosting (Freund & Schapire, 1996)) would allow the system to attribute a label “good”/“not good” to a textual description of a situation of moral dilemma which was not contained in the training corpus is a place for argument.

More closer to the moral core is the 3rd layer, which can be understood as “the layer of rules”. To simplify the understanding: while layer 4 - understood as “the layer of associations amongst data” - can be compared to an anglo-saxon legal system where a decision is based on the precedent, i.e. the first decision of a judge in a case sharing analogic features to a case under study; the activity of layer 3 can be compared to that of a continental judge whose decisions are simple outputs of more general rules induced from exhaustive sets of previous experiences.

Thus, the correct understanding of “moral induction” seems to be crucial in order to implement the robust solution for layer 3 and an inspiration coming from much better studied domain of “unsupervised grammar induction” (Solan, Horn, Ruppin, & Edelman, 2005) may yield encouraging results.

It is not unreasonable to imagine that by applying the induction principles not upon the data, but upon the very rules which were themselves induced, the process would finally converge at the point of some-kind of meta-rule, possibly similar in meaning to that what Kant called “categorical imperative” (Kant, 1785). The advantage of such a “meta-rule” is not only that it is quite easy to implement from programmer’s point of view - in its essence it is nothing else than just an infinite while() loop generating “the representations of possible worlds” and throwing exceptions if ever an “internally inconsistent” world is generated - but that it can be used as a sort of boolean rule of thumb there, where fuzzy thresholds of layers 4 & 3 are unable to supply any decisive result.

The disadvantage of layer 2 is that sometimes it may happen that it shall demand infinite amount of time in order to return the result (A. Turing, 1937). That is far too much especially in the cases where an artificial agent could harm its modified environment by its otherwise harmless activity - imagine, for example, an autonomous transporting agent similar to a car whose circuits got stuck in a while loop after it had hasardously entered the pedestrian zone. For such cases, low-level implementation of fast & frugal harm-reductive inhibitory mechanisms is of utmost importance.

In order to stay consistent with the Tradition, we propose Asimov’s Laws of Robotics (Anderson, 2008) as a base for such mechanisms.

Finally, it is worth to be stated that while layers 4 & 3 are dynamic in their nature, i.e. can be rewritten by inflow of new stimuli from environment, layers 2 & 1 can be embedded into very chips of an artificial agent and could not be modified or disabled without tampering with agent’s hardware.

Believing that such a combination of “two static” and “two dynamic” pillars is in certain sense analogic to a “nature” (i.e. innate) & “nurture” (i.e. acquired) components attributed to the moral faculty of a healthy human being, it may be finally stated that the question which is labeled hereby as a “the central problem of roboethics” is, mutatis mutandi, nothing else than just a postmodern variation upon a much more ancient theme: “How does a parent transform a crying child into an autonomous human being ?”

## References

- Anderson, S. L. (2008). Asimov’s “three laws of robotics” and machine metaethics. *AI & Society*, 22(4), 477-493.
- Andorno, R. (2004). The Precautionary Principle: A New Legal Standard for a Technological Age. *Journal of International Biotechnology Law*, 1(1), 11-19. doi: 10.1515/jibl.2004.1.1.11.
- Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-* (pp. 148-156). Citeseer.
- Kant, I. (1785). *Groundwork of the Metaphysics of Morals*. First published.
- Kurzweil, R. (2000). The age of spiritual machines: When computers exceed human intelligence.
- Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, 11(4), 143-152.
- Searle, J. (1982). The Chinese room revisited. *Behavioral and Brain Sciences*. Retrieved March 11, 2011, from [http://journals.cambridge.org/abstract\\_S0140525X00012425](http://journals.cambridge.org/abstract_S0140525X00012425).
- Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, 102(33), 11629.
- Turing, A. M. (2008). Computing machinery and intelligence. *Parsing the Turing Test*, 23-65.
- Turing, A. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical ...*. Retrieved March 11, 2011, from <http://plms.oxfordjournals.org/content/s2-42/1/230.full.pdf>.
- Vinge, V. (1993). Technological singularity. *VISION-21 Symposium sponsored by NASA Lewis* .

## AFFECTING THE WORLD OR AFFECTING THE MIND?

### *The Role of Mind in Computer Ethics*

JOHNNY HARTZ SØRAKER

*Department of Philosophy, University of Twente*

*j.h.soraker@utwente.nl*

**Abstract:** The purpose of this paper is to draw a distinction between two interrelated yet fundamentally different ways of approaching problems in computer ethics, with the goal of clarifying which problems call for which approaches. In a nutshell, I will draw a distinction between approaches and topics that are primarily concerned with how technologies affect the *world*, on the one hand, and those primarily concerned with how technologies affect our *mind*, on the other. I will argue that the type of approach we choose should be determined on the basis of which of these concerns we are primarily trying to address, which will also shed light on the advantages and disadvantages of the multitude of approaches to be found in ethics of technology. In order to clarify and justify this distinction, I will categorize some common approaches in computer ethics correspondingly, and I will conclude by offering a set of suggestions for how they can and should complement each other in a way that yields an exhaustive analysis of the problem at hand.

The purpose of this paper is to draw a distinction between two interrelated yet fundamentally different ways of approaching problems in computer ethics, with the goal of clarifying which problems call for which approaches. In a nutshell, I will draw a distinction between approaches and topics that are primarily concerned with how technologies affect the *world*, on the one hand, and those primarily concerned with how technologies affect our *mind*, on the other.<sup>13</sup> It should be emphasized at the outset that these categories are not absolute or mutually exclusive – and it is certainly not my intention to argue that one is better than the other. My more modest intention is to argue that the type of approach we choose should be determined on the basis of which of these concerns we are primarily trying to address, which will also shed light on the advantages and disadvantages of the multitude of approaches to be found in ethics of technology.

---

<sup>13</sup> This distinction is reminiscent of Floridi & Sanders' emphasis on the distinction between agent-oriented and patient-oriented ethics (2002), but this distinction is somewhat misleading in this context, because both technology and the mind can have a role as both agent and patient, being both source and target of good and evil.

There is little doubt that technologies affect both the world and the mind, and there is little doubt that there is no sharp distinction between the two. What affects the world can affect the mind, and what affects minds can affect the world – and technology often mediates *between* world and mind. As such, the distinction I am concerned with must necessarily be more of the ‘family resemblance’-type. Still, we can to some degree separate between different ways of assessing these effects, and given the multitude of ethical theories and applied frameworks that are being used in ethics of technology, it is important to be clear about which approach is best suited for which area.

The clearest example of this is probably the distinction between accountability and responsibility. If the purpose of our analysis is to understand what is *accountable* for a given situation, we *can* do this entirely in terms of analyzing changes to the world. After all, an inquiry into accountability is largely an inquiry into causality; what was the *source* of this good or evil (cf. Floridi & Sanders, 2004, p. 371). This also highlights the advantage of using a “mind-less” notion of accountability in cases where (higher-order) mental processes are either non-existent (e.g. artificial agents) or intrinsically distributed (e.g. organizations). If the purpose of our analysis is to understand *responsibility*, however, we are immediately required to include the mind in a much more integral manner. After all, an inquiry into responsibility is an inquiry into such mental terms as *intentions*, *negligence*, and *culpability*. To give another example, when evaluating how Information and Communication Technologies (ICTs) affect privacy, we can focus on how ICTs affect the *world* in a manner that is relevant to privacy, or how it affects our *mind* in a way that is relevant to privacy. The former involves such question as “How do ICTs affect the flow of information”, or what Floridi refers to as ‘ontological friction’ (2005). The latter involves questions such as “How do ICTs affect our expectations about privacy?” and “How can loss of privacy affect our well-being?”. If we look to environmental ethics, we can make a similar distinction between the effects a technological innovation may have on the environment, on the one hand, and their effect on e.g. opinions about sustainability, on the other. We can make a similar distinction when evaluating cultural consequences, by either looking at how technologies may change the material conditions necessary for certain cultural practices, or how they more directly change people’s cultural values and attitudes.

Clearly, the questions are interrelated and both sets of questions should be sought answered in a comprehensive analysis, but the approaches and methods we utilize in doing so will typically be centered on one of the two sets. To clarify this further, we can attempt to categorize different approaches according to their main concerns.

On the one hand, some theories and approaches are particularly good at evaluating how technologies affect the world. Again, one clear example is Floridi’s notion of ‘re-ontologization’ (2005) and the use of an informational level of abstraction, which is an interesting and often insightful way of conceptualizing how the world changes as a result of our increased ability to digitize information . Other examples of this type of approach is Actor-Network theory (Latour, 2005), as well as recent post-phenomenological work on technological mediation (Verbeek, 2005). The strength of these theories is that they shed light on how technologies affect the world and our ways of interacting with the world. They do not, however, say much about how technologies affect the mind. Surely, the changes to the world that they disclose will very often lead to changes in mind, but this is not their main concern.

On the other hand, some theories and approaches are particularly good at evaluating how technologies affect the mind. Among the approaches in this category, we can include approaches that are grounded in some version of virtue ethics or utilitarianism, as well as axiological approaches. The main concern of these approaches is not to understand how technologies affect the world, but rather how they affect our moral character, behavioral dispositions, expectations, quality of life, and so forth. Certainly, technologies often affect our mind *through* changing the world – indeed, they *always* do so if we regard the technology itself as a change to the world. Nevertheless, the main concern of these approaches is not to get a better understanding of how states of affairs in the world change, but rather to get a better understanding of how mental processes change. This is the ultimate *goal* of the analysis. If we take video game violence as an example, a virtue ethical analysis of this phenomenon would not be particularly interested in how these games may affect the physical world, but rather how they will affect the mind of those who interact with them. Will they make them more aggressive, less altruistic, more happy?

One reason for distinguishing between these approaches is that they give rise to different types of normativity, and to show how these can be related to each other. Approaches that are primarily interested in changes to the world can be described as *cautionary*. That is, the effects that technologies have on the world will in many cases imply a *caution*; technology x will lead to change y, and this change might be ethically problematic. In order to take that last step, however, we need approaches that include the mind in order to argue that change y is ethically problematic because it affects the mind in a particular way. This can be seen clearly when teaching computer ethics to pragmatically oriented computer scientists, where showing that technologies change the world will often lead to the perfectly rational question: “That might very well be true, but why is that a problem?”. Answering *that* question must somehow include the mind.

In the full paper, I will further clarify the nature of this distinction, knowing very well that it is problematic and rests on a number of philosophically controversial presuppositions. I will also justify why the mind is essential for most topics in computer ethics, and discuss what this means for how we ought to approach these topics. Some of the main conclusions will be that computer ethics is necessarily and intrinsically a pluralist area of investigation, one that *needs* to address both the world and the mind. More substantially, it will be argued that we need to get a much better understanding of how different approaches can complement each other and how analyses of changes to the world can be integrated into analyses of changes to the mind. I will conclude the paper by offering a few suggestions on how to do so, using privacy as one of the main examples.

## References:

- Floridi, L. (2005). The ontological interpretation of informational privacy. *Ethics and Information Technology*, 7(4), 185-200.
- Floridi, L., & Sanders, J. W. (2002). Mapping the foundationalist debate in computer ethics. *Ethics and Information Technology* 4, 1-9.
- Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, 14(3), 349-379.

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Verbeek, P.-P. (2005). *What things do: philosophical reflections on technology, agency, and design*. University Park, PA: Pennsylvania State University Press.

## THE ETHICS OF AUTOMATED WARFARE

RYAN TONKENS  
York University  
Toronto, Canada  
tonkens@yorku.ca

Autonomous machines of varying degrees are moving onto the battlefield at an overwhelming pace. If left unchallenged, there is good reason to believe that both their level of autonomy and overall sophistication will increase exponentially in the future. In light of this, it is important that we determine whether or not these sorts of robots *should* have a place in warfare.

Here I ask whether the development and use of autonomous military robots is consistent with the tenets of Just War Theory (hereafter JWT).<sup>14</sup> Specifically, the aim of this paper is to offer an in depth (albeit preliminary) analysis of whether the creation and deployment of autonomous machines in military contexts is morally acceptable, by way of assessing the overall justness of automated warfare. *If* automated warfare is unjust, then creating and using robots for this purpose is morally problematic.

The most anticipated application of advanced autonomous machines is in the military sector. Indeed, a disproportionate amount of funding for research on machine autonomy has come from military sources for military applications. Insofar as autonomous robots can perform actions that have serious ethical consequences (in the context of warfare, at least), then they need to be programmed to behave ethically, i.e. to perform only those actions that are in line with the appropriate regulations and agreed upon customs of just war. Contemporary JWT is the received view on how warfare should be conducted. We demand that all (human) combatants abide by the tenets of JWT. Moreover, we expect proper restitution, and go to great lengths to ensure that all breaches of JWT in practice are punished accordingly. If we want to involve autonomous machines in warfare, then they will need to abide by JWT as well.

In this paper I take up four issues towards this end: (1) issues of moral responsibility; (2) discrimination and proportionality; (3) whether the creation of autonomous military machines is consistent with *jus ad bellum* and wider social justice; and (4) whether military machines could be more moral than humans.

(1) JWT demands that someone be morally responsible for actions in war. Given a certain advanced level of machine autonomy, robots will need to be held responsible for their own actions. However, doing so seems futile since they have no capacity to suffer (Sparrow 2007). One potential limitation of Sparrow's analysis, however, is that the range of autonomous machines whereby something (someone) could still be held

---

<sup>14</sup> Just War Theory works in tandem with the international laws of war and rules of engagement as the moral and legal regulations of warfare. Due to space restrictions, I cannot attend to all three herein, so I focus exclusively on JWT.

responsible is quite large. Limiting the autonomy of machines to the point where a human remains in the decision-making/execution loop avoids this problem, since human users are the sort of being that can be punished for their moral wrongdoings.

(2) Autonomous robots will need to be able to accurately and reliably discriminate between legitimate and illegitimate targets (i.e. between combatants and noncombatants, between surrendering combatants and aggressive combatants, between allies and enemies). Whether or not autonomous military machines could be designed to do so in real world military contexts remains an open question, although designing a robot with these abilities does not seem impossible in principle. Regardless, one point that seems uncontroversial is that the level of autonomy and the ability of machines to act in real world contexts will increase much sooner than our ability to perfect their ability to exert the intricacies of discrimination and proportionality at acceptable levels. This is important to recognize because, until autonomous robots can accurately and reliably discriminate between legitimate and illegitimate targets, then they do not meet this requirement of JWT.

(3) If automated warfare fuels widespread social injustice, including injustices outside of the context of warfare specifically, then it is inconsistent with the principles underlying JWT (e.g. justice, fairness, respect). This could manifest itself in many ways, including increasing the likelihood of (unjust) war<sup>15</sup>, decreasing the likelihood of terminating (unjust) war once it had begun, exacerbating gaps between rich and poor nations and strong and weak military forces, *et cetera*. Moreover, the billions of dollars going into the automated military sector could be redirected towards the healthcare or education systems (for example), which could serve to remedy the existing status quo that finds humans of low socioeconomic status with poorer health and lower education, itself a symptom of and catalyst for widespread social injustice.

(4) Despite the possibility that machines could in some sense be more moral than human soldiers under certain circumstances (Arkin 2009; Sullins 2010), automated warfare will also witness its fair share of unethical activity. Although substituting human combatants for machines is appealing in certain ways, automated war would not be less unjust than human warfare overall.

We seem to be seeking to develop autonomous military machines (in part) because we believe that we can treat them like servants and subordinates, yet we also expect them to be military and ethical ‘superiors’. The only way we can bring this about in a morally justifiable manner is if we restrict their sophistication to a point well before they are fully autonomous moral agents (especially akin to *human* moral agents), and hence keep them at a level where we need to keep a human in the loop. But doing so entails continuing to sacrifice human lives in battle, and continuing to endure human moral transgressions and imperfections in decision-making, all *in addition to* the new ethical challenges that accompany automated warfare.

There is good reason to suggest that the creation and use of autonomous military machines is inconsistent with JWT in several respects. This is an important finding. For one thing, it makes it apparent that the creation of certain kinds of autonomous military machines is inconsistent with the moral framework that these robots will be expected to follow. More importantly perhaps, it places the burden of proof on those who want to support the move towards automated warfare and to develop these sorts of machines to demonstrate that they can do so in a morally sustainable (just) manner. Minimizing the level of sophistication of these robots and keeping humans in the military loop seems to

---

<sup>15</sup> McMahan (2009) has argued convincingly that, for diverse and complicated reasons, the majority of wars fought are unjust.

be the most prudent course to adopt, one certainly more palatable than automated warfare *tout court*, although needless to say infinitely less desirable than peace.

## References

- Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots*. Dordrecht: Chapman & Hall.
- Asaro, P. (2008). How just could a robot war be?. In: P. Brey, A. Briggle and K. Waelbers (Eds.), *Current Issues in Computing and Philosophy* (pp.50-64). Amsterdam: IOS Press.
- Guarini, M. & Bello, P. (forthcoming). Robotic warfare: Some challenges in moving from non-civilian to civilian theaters. In: P. Lin, G. Bekey and K. Abney (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge: MIT Press.
- McMahan, J. (2009). *Killing in War*. Oxford: Clarendon Press.
- Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, 24 (1), 62-77.
- Sullins, J. (2010). RoboWarfare: Can robots be more ethical than humans on the battlefield? *Journal of Ethics and Information Technology*, 12: 263-275.

## CAREBOTS AND CAREGIVERS

### *Robotics and the Ethical Ideal of Care*

SHANNON VALLOR

*Department of Philosophy, Santa Clara University*

*500 El Camino Real*

*Santa Clara, CA 95053 USA*

**Abstract.** In the 21st century we stand on the threshold of welcoming robots into domains of human activity that will expand their presence in our lives dramatically. One provocative new frontier in robotics, driven by a convergence of demographic, economic, cultural and institutional pressures, is the development of ‘carebots’ - robots intended to assist or replace human caregivers in the practice of caring for vulnerable persons such as the elderly, young, sick or disabled. I argue that existing reflections on the ethical implications of carebots have thus far neglected a critical dimension of the issue: namely, the potential moral value of caregiving practices for caregivers. Instead, the scholarly dialogue has largely focused on the potential benefits and risks to care recipients. Where caregivers have been explicitly considered, it is strictly in terms of how they might benefit from having the burdens of care reduced by carebots. I stipulate here that properly designed and implemented carebots might improve the lives of cared-for and caregivers in ways that would be ethically desirable. Given the grave deficiencies of existing social mechanisms for supporting caregivers, their use may even be ethically obligatory in the absence of acceptable alternatives. Yet I argue that we ought to forestall such judgments until we have first adequately reflected upon the existence of goods internal to the practice of caregiving that we might not wish to surrender, or that it might be unwise to surrender even if we might often wish to do so. Such reflection, I claim, gives rise to considerations that must be weighed alongside the likely impact of carebots on care recipients. In order to initiate such reflection, I examine the goods internal to caring practices and the potential impact of carebots on caregivers by means of three complementary ethical approaches: virtue ethics, care ethics and the capabilities approach. I show that each of these frameworks can be used to shed light on the contexts in which carebots might deprive potential caregivers of important moral goods central to caring practices, as well as those contexts in which carebots might help caregivers sustain or even enrich those practices.

## **1. Introduction**

We stand on the threshold of welcoming robots into domains of human activity that will expand their presence in our lives dramatically. One provocative new frontier is the development of ‘carebots’ - robots intended to assist or replace human caregivers in the practice of caring for vulnerable persons such as the elderly, young, sick or disabled. Yet existing philosophical reflections on the ethical implications of carebots have thus far neglected a critical dimension of the issue: the potential moral value of caregiving practices for caregivers. Instead, the dialogue has largely focused on the potential benefits and risks to care recipients. Indeed, properly designed and implemented carebots might improve the lives of both cared-for and caregivers in ways that would be ethically desirable. Their use may even be ethically obligatory in the absence of acceptable alternatives. Yet I argue that such judgments are premature until we have adequately reflected upon the potential existence of goods internal to the practice of caregiving that we might not wish to surrender, or that it might be unwise to surrender even if we might often wish to do so.

Such reflection, I claim, gives rise to considerations that must be weighed alongside considerations of the likely impact of carebots on care recipients. Taking as a guiding insight Coeckelbergh’s (2009) claim that we must look beyond mere application of “external” ethical criteria for human-robot relations, I propose to examine the goods internal to caring practices and the potential impact of carebots on caregivers by means of three complementary ethical approaches: virtue ethics, care ethics and the capabilities approach. Each of these philosophical frameworks sheds new light on: 1) the contexts in which carebots might deprive potential caregivers of important moral goods central to caring practices, 2) contexts in which carebots might help caregivers sustain or even enrich those practices, and 3) the specific nature of those moral goods.

## **2. Carebots and the ethical significance of caring practices**

### **2.1. THE VIRTUES OF CARE**

A virtue-ethical account offers rich resources for our inquiry in the form of a range of moral virtues that can be cultivated and sustained through caring practices. Patience, understanding, charity, prudence, reciprocity and empathy can each be cultivated through sustained caring activity. ‘Excellent carers’ manifest a powerful ability to anticipate and interpret the needs of others, even when not explicitly communicated. They habitually express effective responses to those needs, even in unusual or rapidly changing situations. They are able to maintain emotional bonds with others, even under physically and mentally demanding circumstances. They enable the autonomy and self-expression of those they care for, to whatever degree possible. If Aristotle is right that the virtues must be cultivated by habitual performance of practices appropriate to their expression (1984, 1103b1), then caring practices are an important, perhaps even essential, part of one’s moral development. This is a compelling reason to examine the potential impact of carebots designed to free us from those practices. Yet carebots have also been proposed as a means of facilitating deeper human engagement in caring practices, by taking over routine or unpleasant chores that drain our energy for giving

good care. (Coeckelbergh, 2010). This suggests the need for a sustained study of which kinds of caring practices are most critical for the cultivation of caring virtues. Such a study, guided by a virtue-ethical framework, could greatly assist the ethical implementation of carebots by providing carebot developers, institutions, and caregivers with critical information about the moral value of various caregiving practices.

## 2.2. CARE ETHICS, CAREBOTS AND THE ETHICAL IDEAL

Care ethics provides another source of insight. Noddings (1984) offers an account of the ‘caring relation’ that takes it to be ethically primary in human existence - a source not only of individual virtues, but also (and more fundamentally) of an ethical ideal that motivates and guides human flourishing. I will argue that carebots might be used to modify contexts of care in ways that preserve or enhance this ethical ideal, allowing us to be engrossed in the needs of the other, moved to attend to them, and open to the responses of those for whom we care. Yet Noddings’ account can also remind us that our aim is not to be liberated from the caring relation itself, for if she is right, this is the only human relation through which our own ethical ideal can be nurtured.

## 2.3. CARING AND THE CAPABILITIES APPROACH

Nussbaum’s capabilities approach provides a third perspective on the goods internal to caring practices. Among the capabilities emphasized by Nussbaum as critical to human flourishing (2006, 76-77), I argue that affiliation, practical reason and emotion are each realized, to a critical degree, through caring practices. For it is at least partly through providing care that I develop the intimate knowledge of human vulnerability needed to fully exercise these capabilities. We must therefore reflect carefully on the way in which the introduction of carebots in society could inhibit or enhance their development.

## 5. Conclusion

Together these conceptual frameworks can remind us that in reflecting upon the ethical portent of carebot technology, we must consider more than just the quality of care robots can give, the relevant preferences and likely reactions of cared-for, or the strong social pressures we face to better meet the needs of the vulnerable among us. These are all serious ethical considerations to which we must carefully attend in weighing the costs, benefits and risks of carebot implementation – but it is of critical importance that we not overlook the moral goods internal to caring itself.

## References

- Aristotle (1984). *The Complete Works of Aristotle: Revised Oxford Translation*. Princeton: Princeton University Press.
- Coeckelbergh, M. (2009). Personal robots, appearance and human good: A methodological reflection on roboethics. *International Journal of Roboethics*, 1(3), 217-221.
- Coeckelbergh, M. (2010). Health care, capabilities and AI assistive technologies. *Ethical Theory and Moral Practice*, 13(2), 181-190.

Noddings, N. (1984). *Caring: A Feminine Approach to Ethics and Moral Education*. Berkeley: UC Press.

Nussbaum, M. (2006). *Frontiers of Justice: Disability, Nationality, Species Membership*. Cambridge: Harvard University Press.

## CO-CONSTRUCTION AND CO-MANAGEMENT OF ONLINE IDENTITIES

### *A Confucian Perspective*

PAK-HANG WONG  
*Department of Philosophy,  
University of Twente*

**Abstract.** In information and computer ethics, the discussion of personal identities online (PIOs) is often framed as if individuals are victims who need protection, e.g. privacy, identity theft, etc. In this respect, many of the discussions related to PIOs in the current literature are negative in that they aim to provide and justify certain constraint and restrictions on (the use of) PIOs. While the issues concerning privacy, identity theft, etc. are undoubtedly important, the lone focus on negative aspects related to PIOs is undesirable, for it has undermined the scope of issues related to PIOs, particularly, the more positive issues pertaining to PIOs, e.g. how we should construct and manage our PIOs. Recently, Noëmi Manders-Huits has studied the notion of “identity management” in the context of information technology. Manders-Huits’s article is significant, because she has explicitly turned away from the negative issues and moved on to issues about the construction and management of identities in IT, which are far more positive. As such, her discussion introduced a new area of research that is so far largely neglected. Although her study of identity management is illuminating, I think her account is unsatisfactory ultimately, as she failed to properly acknowledge one important facet of PIOs, namely they are co-constructed and co-managed. The aim of this paper, therefore, is to remind of the fact that PIOs are co-constructed and co-managed, and to identify some conceptual and ethical issues arise from it. Finally, I will outline the answers to the issues using a Confucian notion of personhood and identity.

### 1.

In information and computer ethics, the discussion of personal identities online (PIOs) is often framed as if individuals are victims who need protection, e.g. privacy, identity theft, etc. In this respect, many of the discussions related to PIOs in the current literature are negative in that they aim to provide and justify certain constraint and restrictions on (the use of) PIOs. As Shoemaker noted, most of the literature in the field attempted to specify “a protected zone of private information, consisting in information about me.” (Shoemaker 2010, 3-4) While the issues concerning privacy, identity theft,

etc. are undoubtedly important, the lone focus on negative aspects related to PIOs is undesirable, for it has undermined the scope of issues related to PIOs, particularly, the more positive issues pertaining to PIOs, e.g. how we should construct and manage our PIOs. Recently, Noëmi Manders-Huits (2010) has studied the notion of “identity management” in the context of information technology. Manders-Huits’s article is significant, because she has explicitly turned away from the negative issues and moved on to issues about the construction and management of identities in IT, which are far more positive. As such, her discussion introduced a new area of research that is so far largely neglected. Although her study of identity management is illuminating, I think her account of is unsatisfactory ultimately, as she failed to properly acknowledge one important facet of online identities, namely online identities are co-constructed and co-managed. The aim of this paper, therefore, is to remind of the fact that online identities are co-constructed and co-managed, and to identify the conceptual and ethical issues arise from it. Finally, I will outline the answers to the issues using a Confucian notion of personhood and identity.

I will begin this paper with Manders-Huits’s account of identity management. According to Manders-Huits, there are two senses of “identity management”. The first is being used predominantly in the technical discourse, where identity management refers to the practice of collecting, organising and, subsequently, utilising personal information for the purpose of (re-)identification and categorisation. (Manders-Huits 2010, 47) And, the second sense of identity management involves not only a set of description about the individual; it also involves reflexive, self-identification with some sets of beliefs, values or ideals, where those beliefs, values and/or ideals provide reasons for our actions and, at the same time, make the actions genuinely ours. (see, e.g. Korsgaard 1996; Frankfurt 1998, 1999, 2004 & 2006) Identity management in the second sense, therefore, requires individuals to manage their beliefs, values and ideals, and to resolve possible conflicts among them. (Manders-Huits 2010, 48-9) As she rightly pointed out, identity management is an issue deserving more attention, as there is a discrepancy between the two senses of “identity management”, and the moral and practical dimension of identity is currently not being taken into account in both the technical discourse and in the technologies. Yet, for the centrality of moral and practical identity in our lives, the negligence of it has to be rectified. I agree entirely with her claim, but I shall also point out that identity management will become more important as information technology continues to develop and being adopted.

## 2.

As information technology (and the Web) continues to advance, it will – to use Luciano Floridi’s terminology – re-ontologise the nature of ourselves and our world. According to Floridi, we are (becoming) inforgs, i.e. “connected informational organisms” living in an infosphere, i.e. “an environment constituted by all informational entities [...], their properties, interactions, processes and mutual relations.” Floridi (2007, 60, 62 & 59) At certain point, Floridi argued, the boundaries between the life offline and the life online will eventually evaporate, and by then individuals will be living in the Web Onlife. Among other characteristics, the onlife of inforgs in an infosphere is characterised by instant, seamless exchanges of offline and online information. In other words, the flow

of (personal) information will become, at least, bi-directional. What it means is that when individuals act on the Web, it will have immediate and direct impacts on their non-Web counterparts. In this scenario, identity management for online identities becomes essential. Since it will no longer be possible to distinguish the offline and the online, it will be impossible to dissociate online identities from offline identities too. Or, to put it differently, what remains are onlife identities.

While Manders-Huits is right to point out that identity management is an important issue for researchers in information and computer ethics, I shall argue that her account of identity management is unsatisfactory, because she has failed to properly acknowledge the fact that online identities are co-constructed and co-managed by multiple parties. This failure is reflected in her suggestion to engineers and technology designers, when she remarked that they “should provide ways for individuals to construct and maintain their [reflexive, self-identification with some sets of beliefs, values or ideals] and [some sets of descriptions about themselves], in addition to their administrative, forensic counterpart.” (Manders-Huits 2010, 54) It is obvious that the emphasis is on empowering individuals in managing their personal information. Yet, what is missing here is that: while it is true that individuals construct and manage their online identities, we are not the only one who contributes to their construction and management. For example, a person’s profile on Facebook is not only what that person inputs, but the totality of information on the profile, including his/her friends, conversations, etc. In other words, not all identity-related information is under the person’s control. In light of this, I shall argue that there is a need to reconceptualise PIOs in terms of co-construction and co-management; and, I shall also argue that unless the person is omnipotent and omnipresence, empowering individuals is always insufficient.

### 3.

At this point, I suggest that we can learn a lesson from Confucianism. I will point out that Confucians conceptualised personhood and identity as inherently interdependent and relational. (Wong 2004; Lai 2006; Yu & Fan 2007) And the Confucian personhood and identity, I shall argue, provide us an alternative way to conceptualise PIOs, which can take into account the co-construction and co-management of PIOs. Moreover, accompanied with the Confucian personhood and identity is an ethics, which is based on individuals’ social roles. (Nuyen 2009) Here, I will suggest that the role-based ethics in Confucianism offers a fitting complement to the Manders-Huits’s strategy of individual empowerment.

### References

- Floridi, L. (2007). A look into the Future Impact of ICT on Our Lives. *The Information Society*, 23 (1), 59-64.
- Floridi, L. (2009). The Semantic Web vs. Web 2.0: A Philosophical Assessment. *Episteme*, 6, 25-37.
- Frankfurt, H. (1988). *The importance of what we care about: philosophical essays*. Cambridge: Cambridge University Press

- Frankfurt, H. (1999). *Necessity, volition, and love*. Cambridge: Cambridge University Press
- Frankfurt, H. (2004). *The reasons of love*. Princeton, N.J.: Princeton University Press.
- Frankfurt, H. (2006) *Taking ourselves seriously and getting it right*. Stanford, Calif.: Stanford University Press.
- Korsgaard, C. (1996). *The sources of normativity*. Cambridge: Cambridge University Press.
- Lai, Karyn (2006). *Learning from Chinese Philosophies: Ethics of Interdependent and Contextualised Self*. UK: Ashgate
- Manders-Huits, N. (2010). Practical versus moral identities in identity management. *Ethics and Information Technology*, 12 (1), 43-55.
- Nuyen, A.T. (2009) Moral Obligation and Moral Motivation in Confucian Role-Based Ethics. *Dao*, 8, 1-11
- Shoemaker, D. W. (2010). Self-exposure and exposure of the self: information privacy and the presentation of identity. *Ethics and Information Technology*, 12 (1), 3-15.
- Tavani, H. T. (2008). Informational Privacy: Concepts, Theories, and Controversies. In K.E. Himma and H.T. Tavani (Eds.), *The Handbook of Information and Computer Ethics* (pp. 131-164). Hoboken, NJ: John Wiley and Sons.
- Wong, David (2004) Relational and Autonomous Selves. *Journal of Chinese Philosophy*, 31 (4), 419-432
- Yu, Erika & Fan, Ruiping. (2007) A Confucian View of Personhood and Bioethics. *Bioethical Inquiry*, 4, 171-179

# **Track VI: Multidisciplinary Perspectives**

## REFLECTIVE INEQUILIBRIUM

BERT BAUMGAERTNER  
*University of California, Davis*  
*1240 Social Sciences and Humanities*  
*University of California, Davis*  
*One Shields Avenue*  
*Davis, CA 95616*

**Abstract.** I show that under a traditional introspective method of philosophical investigation, certain projects of conceptual analysis are bounded by a reflective *inequilibrium*. That is, although it is possible to make some progress towards bringing our classificatory intuitions and the relevant criteria into agreement, there is a barrier that cannot be overcome with traditional methods when the concept in question is plastic. We can show the limitations of the traditional method of conceptual analysis by considering its computational analog. Suppose we have an algorithm *C* that determines a set of cases that fall under a given concept and another algorithm *T* which tests cases by consulting *C* (which responds with 'Yes' or 'No'). If *C* is static (and decidable), then in principle *T* can develop a criterion for it. Moreover, every verification procedure that *T* uses to check the match yields consistent results. However, this turns out not to be the case when *C* is plastic. Even if we assume the best case scenario in which a proposed criterion matches the set of cases determined by the concept, testing cases near the boundary moves the boundary, and so the criterion will no longer match. So even if an algorithm gets a match via a lucky guess, it is unable to verify the match. A state of affairs where no perfect match can be verified is a *reflective inequilibrium*. That some concepts are plastic is supported by empirical evidence which shows that classificatory intuitions can be affected by the order in which cases are considered. Swain et al. (2008) found that individual intuitions can vary according to whether, and which, other thought experiments were considered first. It is likely that the varying intuitions track shifts in the classificatory dispositions of our concepts. In fact, it is well accepted in cognitive psychology and cognitive science that human concepts are flexible and dynamic in this way. Interestingly then, a computational approach to traditional introspect methodology thereby gives us a possible explanation for why conceptual analysis is so difficult and usually unsuccessful.

### Extended Abstract

In this paper, I show the far-reaching effects of the computational turn by shedding light on a traditional problem. Specifically, I show that under a traditional introspective method of philosophical investigation, certain projects of conceptual analysis are bounded by a reflective *inequilibrium*.

In the philosophical literature, particularly in certain domains of epistemology, it is assumed that a conceptual analysis of knowledge, for example, is possible through a process of reflective equilibrium. This process is a virtuous circle, where we make some headway on settling which cases count as knowledge in order to develop some criteria, and we let the development of criteria help us settle on which cases count as knowledge. As I will show however, although it is possible to make some progress towards bringing these two into agreement, there is a barrier that cannot be overcome with traditional methods when the concept in question is plastic. Since it is plausible that our concept of knowledge is plastic (Weinberg et al., 2001), the possible progress of an analysis given traditional methods is bounded by a reflective inequilibrium.

More specifically, a traditional method of doing conceptual analysis can be characterized as the attempt to bring into agreement our classificatory intuitions about cases and a proposed criterion that defines the relevant set of cases. We then proceed by testing proposed criteria. This is done by a) introspectively checking whether every possible case as specified by a criterion is an instance of the concept in question, and b) introspectively checking whether every possible instance of the concept in question is a possible case specified by the criterion.

We can show the limitations of the traditional method of conceptual analysis by considering its computational analog. We have an algorithm *C* that determines a set of cases that fall under a given concept. We then have another algorithm *T* which tests cases by consulting *C* (which responds with 'Yes' or 'No'). Given data from *C*, *T* attempts to develop a criterion for the set of cases determined by *C*. If this set is static (and decidable), then in principle *T* can develop a criterion for it. Moreover, every verification procedure that *T* uses to check the match yields consistent results. However, this turns out not to be the case when *C* is plastic.

Let us assume the best case scenario in which a proposed criterion matches *C*. In order for *T* to verify the match, it must test some cases again. But since *C* is plastic, testing cases near the boundary moves the boundary, and so the criterion will no longer match *C*. Then *T* will get an inconsistent result for some verification procedure. So even if *T* gets a match via a lucky guess, it is unable to verify the match. Let us call a state of affairs where no perfect match can be verified a *reflective inequilibrium*.

We have appealed to an intuitive notion of plasticity. More rigorously, plasticity can be implemented in an artificial cognitive system by the specification of two features: i) the conditions for when the boundary of a concept shifts, and ii) how much the boundary of the concept shifts. Such algorithms behave in the following way. When given cases to classify near the boundary, the boundary shifts by some amount, so that future cases which may have been classified positively (negatively) may now be classified negatively (positively). Boundary shifting is more or less stable depending on how the cases are selected for testing and how features (i) and (ii) are specified.

That some concepts are plastic is supported by empirical evidence which shows that classificatory intuitions can be affected by the order in which cases are considered. For example, Swain et al. (2008) found that individual intuitions can vary according to whether, and which, other thought experiments were considered first. It is natural to suppose that the varying intuitions track shifts in the classificatory dispositions of our concepts. In fact, it is well accepted in cognitive psychology and cognitive science that human concepts are flexible and dynamic in this way. Psychologists such as Laurence Barsalou (1987) and James Hampton (2007) have suggested that this is a good thing, for

it provides us with the capacity to track environmental changes while maintaining the identity of the relevant concept(s). Let the *plasticity hypothesis* be the hypothesis that our concepts are apt to change their classificatory dispositions.

In sum, taking a computational approach to traditional introspective conceptual analysis illuminates the limitations of this particular methodology. It is common to think that a barometer of how well we understand cognitive capacities is our ability at simulating artificial systems. Given that we have adequate algorithmic implementations of the plasticity hypothesis and the traditional methodology, we can rigorously prove limitations of the traditional methodology. We thereby have a possible explanation for why conceptual analysis is so difficult and usually unsuccessful -- introspection can provably only take us part of the way. Consequently, the computational approach can make way for the development of additional tools to study human capacities of categorization.

### **Acknowledgements**

Thanks to Adam Sennet and attendees of the philosophy graduate student workshop at UC Davis for helpful suggestions in the initial development of the ideas. Special thanks to Bernard Molyneux for comments and support.

### **References**

- Barsalou, L. (1987). The instability of graded structure: Implications for the nature of concepts. In: U. Neisser (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization* (pp. 101–140). Cambridge: Cambridge University Press.
- Hampton, J. (2007). Typicality, graded membership, and vagueness *Cognitive Science: A Multidisciplinary Journal* 31 (3) (pp. 355–384).
- Swain, S., J. Alexander, and J. Weinberg (2008). The instability of philosophical intuitions: Running hot and cold on truetemp. *Philosophy and Phenomenological Research* 76 (1) (pp. 138–155).
- Weinberg, J., S. Nichols, and S. Stich (2001). Normativity and epistemic intuitions. *Philosophical Topics* 29 (1-2) (pp. 429–460).

## THE INFORMATION-COMPUTATION TURN: A HACKING-TYPE REVOLUTION

ISRAEL BELFER

*Science, Technology and Society Program,  
Bar Ilan University  
Ramat Gan, Israel*

**Abstract.** Hacking's Styles of Reasoning (Hacking 1981, 1992) are utilized to describe the impact Information Theory has had on science in the 20th century in theory and application. A generalized, *Information-laden* scientific style of reasoning is introduced, generalizing the information-theoretical and computational turn in science and society. Information-laden science will be examined according to Hacking's criteria for a new Style, and its associated 'revolution' (Schweber and Watcher, 2000). These criteria include a new scientific vocabulary as well as a wider social and conceptual context. The specific branch of science chosen to exhibit the new style is physics, which manifests a wide range of a style's attributes: science in an information-age ('e-science'); hard-theoretical physics such as Black-Hole Thermodynamics (BHTD) and the consequent Black-Hole Wars (Suskind, 2008); the advent of Quantum Information Theory (QIT) – namely Quantum Information and Quantum Computation.

### 1. Introduction – Hacking Type Revolutions

Hacking's Styles of Reasoning (Hacking 1982, 1992; Crombie, 1994) are meta-concepts that arrange the scheme of ideas and practices in science and society. They are described as:

“The active promotion and diversification of the scientific methods of late medieval and early modern Europe reflected the general growth of a research mentality in European society, a mentality conditioned and increasingly committed by its circumstances to expect and to look actively for problems to formulate and solve, rather than for an accepted consensus without argument. The varieties of scientific method so brought into play may be distinguished as:

- (a) the simple postulation established in the mathematical sciences,
- (b) the experimental exploration and measurement of more complex observable relations,
- (c) the hypothetical construction of analogical models,
- (d) the ordering of variety by comparison and taxonomy,
- (e) the statistical analysis of regularities of populations and the calculus of probabilities, an

(f) the historical derivation of genetic development.

The first three of these methods concern essentially the science of individual regularities, and the second three the science of the regularities of populations ordered in space and time.”

The rise of a Style of Reasoning manifests in a Hacking-Type Revolution that accompanies a new Style.

#### 1.1 A NEW HACKING-TYPE REVOLUTION

Schweber & Watcher (2000) recognized in the computational (information-processing) revolution the rise of such a Style: “We are witnessing another Hacking type revolution, which for lack of a better name we call the ‘complex systems modeling and simulation’ revolution, for complexity is one of its buzzwords and **mathematical modeling and simulation on computers constitute** its style of reasoning”. This Style and its revolution should be adopted and combined with the ubiquity of Information-Theoretical terminology in science (Arndt, 2004), into a generalized form. That is, a Hacking-Type revolution of Information-laden science, with **digitized Information** as its Style.

By expanding on the same theme of the Hacking type revolution to include communication and cryptography, one achieves more than a parceling together of the theoretical basis for these fields of research. It in fact relays a basic theme in science and technology, since communication and computation – Information transfer and processing – are inextricably linked theoretically and practically. The common thread connecting all of these theoretical approaches and applied technologies is the modern concept of quantified information.

#### 1.2 INFORMATION-LADEN PHYSICS

The technological and theoretical growth embodied in the fields of computation and communication amalgamates into a Style of Reasoning with Digitized Information (Shannon, 1948) at its core: That is, information and its measures (Arndt, 2004). A science laden with Information (paraphrasing ‘theory-laden’ science) is saturated with direct and indirect reliance on IT and Information measures for defining problems and their solutions, influencing the theory and the practice of science. Experiment becomes data acquisitions (Brillouin, 1956); analysis - the computerized simulation and processing of relevant datasets.

Much of this process is due to Maxwell’s Demon (Leff, 2003), the thought experiment that challenged the second law of thermodynamics since the end of the 19<sup>th</sup> century. Attempts to deal with it catalyzed lines of theoretical research that primed physics for a turn towards Information, prompting the tight connection between the thermodynamics of computation and IT (Bennet, 1973).

This shift is reinforced by a deeper moment in abstract theoretical work: IT as scientific modeling of nature, such as the Maximum Entropy Principle (Jayens, 1957). The declaration that ‘Information as physical’ (Landauer, 1991; Karnani et al, 2009) connects communication and computation together with fundamental physics and the second law of thermodynamics. Considered by some as ‘the new language of science’ (von Bayer, 2005), a new ‘metaparadigm’ in popularized depictions of the change (Siegfried, 2000; Seife 2006).

## 2. New Fields of Information-Laden Physics

The 20th century saw the development of core mathematical-physics imbued with IT (von Baeyer, 2005), i.e. *Information-laden science*. Jakob Bekenstein's seminal work on Black Hole Thermodynamics (BHTD) (Bekenstein, 1973,2006). Fields of research such as Quantum Information Theory (Fuchs, 2010) and String Theory (Susskind, 2008) do more than utilize Shannon's Information-Entropy measure. They link physical reality to computation and cryptography.

BHTD and M-Theory produced the Holographic Principle (t'Hooft, 1993; Susskind, 1995) according to which physical reality is encoded onto the surface area of the universe. QIT bodes the possibilities of pan-computationalism (Lloyd, 2006; Feynman, 1981; Zuse, 1967) with all physical phenomena understood as bit-flipping. Wheeler (1990) takes it even further: every physical object essentially Informational – his famous aphorism “It from Bit”.

## 3. New Style – Spheres of Science and Society

### 3.1 NEW SENTENCES, OBJECTS AND LAWS.

A new Style enjoys a new semantic field of definitions, sentences and criteria for the proper conduct of science (Hacking, 1992). The new aforementioned topics and disciplines in science are built on precisely such constructs. It is through Information terminology that the Holographic principle and its ramifications on the criteria for a well-constructed M-Theory can be expressed; that the computational universe can be entertained and weighed as a model for physical reality

### 3.2 THE INFORMATION AGE

The wider social setting for these changes in science are explored in the sociological, economic and political research of the Information-Age (Castells, 2004). The Theoretical, applied scientific and technological aspects of the Information-laden revolution are organic to this social moment.

## Acknowledgements

I would like to thank Prof. Silvan Schweber and Dr. Raz Chen Moris for their great support in all stages of this research. I would also like to thank Dr. Chris Fuchs for the great conversations and discussions (on QIT and Chupakabras).

## References

Arndt, Christoph (2004), *Information Measures: Information and Its Description in Science and Engineering*. Heidelberg-Berlin: Springer.

- von Baeyer, Hans Christian (2005), *Information: The new Language of Science*. Harvard University Press.
- Bekenstein, Jakob (1973), Black Holes and Entropy. *Phys. Rev. D*7, 2333.
- Bekenstein, Jacob (2006). *Of Gravity, Black Holes and Information*. Rome: Di Renzo Editore.
- Bennett C. H. (1973). Logical reversibility of computation. *IBM Journal of Research and Development*, 17(6), 525-532.
- Brillouin, Leon (1956). *Science and Information Theory*. Mineola, N.Y: Dover.
- Castells, Manuel (2004). *Informationalism, Networks, and the Network Society: a Theoretical Blueprinting*. Northampton, MA: Edward Elgar.
- Feynman Richard P. (1981). Simulating Physics with Computer [Keynote speech in 1<sup>st</sup> conference on Physics and Computation, MIT 1981]. *International Journal of Theoretical Physics*, 21(6/7), 467-488, 1982.
- Fuchs Christopher A. (2010). *Coming of Age With Quantum Information: Notes on a Paulian Idea*. Cambridge University Press.
- Hacking, Ian (1981). From the Emergence of Probability to the Erosion of Determinism. in J. Hintikka, D. Gruender and E. Agazzi E (Eds), *Probabilistic Thinking, Thermodynamics and the Interaction of the History and Philosophy of Science, Proceedings of the 1978 Pisa Conference on the History and Philosophy of Science* (Vol. II, pp. 105-123). Dordrecht: Reidel.
- Hacking, Ian (1992). 'Style' for Historians and Philosophers, In *Historical Ontology*, Harvard University Press, 178-200
- Hawking, Steven W. (July 2005), Information Loss in Black Holes, arxiv:hep-th/0507171
- 't Hooft, G (1993), Dimensional Reduction in Quantum Gravity, 1993, arXiv:gr-qc/9310026v2
- Jaynes, Edwin T. (1957), Information Theory and Statistical Mechanics. *Physical Review* 106, 620-630
- Landauer, R. (1991) Information is physical. *Physics Today*, May 1991.
- Leff, Harvey S., Rex, Andrew F. (Eds), *Maxwell's Demon 2: Entropy, Classical and Quantum Information*. CRC Press 2003.
- Lloyd Seth (2006) *Programming The Universe: A Quantum Computer Scientist Takes On the Cosmos*. New York: Random House.
- Schweber S., Watcher M. (2000). Complex Systems, Modelling and Simulation. *Stud. Hist. Phil. Mod. Phys.* 31(4), 583-609.
- Susskind, Leonard (1995). The World as a Hologram. *J.Math.Phys.*36:6377-6396.
- Susskind, Leonard (2008). *The Black Hole War: My battle with Stephen Hawking to make the world safe for quantum mechanics*. Little, Brown and Co.
- Shannon, Claude. E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27, 379-423.
- Wheeler, J. A. (1990). Information, Physics, Quantum: The Search for Links. In W. H. Zureck (Editor) *Complexity, Entropy, and the Physics of Information*. Redwood City, Cal.: Addison Wesley.
- Konrad Zuse (1967). Rechnender Raum. *Elektronische Datenverarbeitung*, 8, 336-344.

## HOW MUCH DO FORMAL NARRATIVE ANNOTATIONS DIFFER?

### *A Proppian Case Study*

RENS BOD

*Institute for Logic, Language and Computation  
Universiteit van Amsterdam*

AND

BENEDIKT LÖWE

*Institute for Logic, Language and Computation  
Universiteit van Amsterdam*

AND

SANCHIT SARAF

*Institute for Logic, Language and Computation  
Universiteit van Amsterdam*

**Abstract.** The formal study of narratives goes back to the Russian structuralist school, paradigmatically represented by the 1928 study *Morphology of the Folktale* by Vladimir Propp. Researchers in the field of computational narratology have developed the general Proppian methodology into various formal and computational frameworks for the analysis, automated understanding and generation of narratives. Methodological issues in this research field give rise to concrete research questions such as “How much does the representation of a narrative in a given formal framework depend on subjective decisions of the formalizer?” touching philosophy of computing and philosophy of information. In order to approach the mentioned question, we consider the process of formal representation of a narrative as a natural analogue of the task of annotation in computational linguistics and corpus linguistics. We use the Russian folktales formalized by Propp and let them be formalized by annotators according to Propp's system, evaluating these results according to the standards of inter-annotator agreement.

The formal study of narratives goes back to the Russian structuralist school, paradigmatically represented by the 1928 study *Morphology of the Folktale* by Vladimir Propp (1928) in which he identifies seven *dramatis personae* and 31 functions that allow him to formally analyse a corpus of Russian folktales.

Researchers in the field of computational narratology (or “computational models of narrative”) have developed the general Proppian methodology into various formal and computational frameworks for the analysis, automated understanding and generation of narratives. Examples for this are Lehnert (1981)'s *Plot Units*, Rumelhart (1980)'s *Story Grammars*, Schank (1982)'s *Thematic Organization Points* (TOPs), Dyer (1983)'s *Thematic Abstraction Units* (TAUs), or Turner (1994)'s *Planning Advice Themes* (PATs). Over the last decades, the main interest of this research community lay in the technical challenges that the computational treatment of narratives brings, but recently, there is again increased interest in the methodological and conceptual issues involved, linking this research closely to questions of the philosophy of information (cf. the paper (Löwe to appear) presented at the *3rd Workshop for the Philosophy of Information*). This interest is witnessed by workshops such as the recent AAAI workshop on *Computational Models of Narrative* that brought researchers from this field together with philosophers, narratologists and professional story tellers. The methodological issues involved give rise to concrete research questions such as

- How do you compare formal frameworks of narrative? (Cf. Löwe 2010 and Löwe to appear.)
- How do you assess the quality of a formal framework of narrative?
- How much does the representation of a narrative in a given formal framework depend on subjective decisions of the formalizer?

Question 1. is a genuinely philosophical question, but also the more technical questions 2. and 3. are very relevant for gaining philosophical insight into what constitutes the formal core of the concept of narrative. In this paper, we approach question 3. of the above list. To this end, we think of the process of formal representation of a narrative in a formal system as a natural analogue of the task of annotation in corpus linguistics and computational linguistics. Whereas typical annotation tasks involve annotation of sentences or discourses (cf., e.g., Marcus et al. 1993, Brants 2000, Passonneau et al. 2006), the formalization or annotation of a narrative is at the next level of complexity, involving sequences or systems of discourses, connected to a narrative. First studies suggest that question 3. is not easy to tackle for the following reasons: First, ambiguity which in typical linguistic annotation is a rather confined phenomenon becomes ubiquitous at the level of narratives: the natural answer to a formalization task is not one annotation, but a family of consistent annotations (cf. Löwe 2010, §2). Secondly, even allowing for multiple annotations, it is not clear whether consensus about whether a given annotation is a valid representation of a narrative is easy to achieve.

Of course, these questions naturally reflect a well-known discussion from computational linguistics: in sentence- or discourse-level annotation, the quality of annotation is typically studied as *inter-annotator agreement* (Carletta et al. 1997, Marcu et al. 1999). For the annotation or formalization of narratives, no such analysis has ever been done, not even with the oldest and best-known formal approach to narrative structure, the Proppian narratemes.

In this study, we use English translations of the Afanas'ev tales formalized by Propp (Afanas'ev 1973), train a group of annotators in the use of Propp's system, and then let them formalize a selection of tales in that formal framework. We evaluate these results according to the standards of inter-annotator agreement from computational and corpus linguistics (Carletta et al. 1997).

## References

- Afanas'ev, A (1973). *Russian fairy tales*. Pantheon. Translation by Norbert Guterman from the collections of Aleksandr Afanasev. Folkloristic commentary by Roman Jakobson.
- Brants, T. (2000). Inter-annotator agreement for a German newspaper corpus. In: *Proceedings Second International Conference on Language Resources and Evaluation LREC-2000*.
- Carletta, J.C., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. & Anderson, A (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13-31.
- Dyer, M.G. (1983). *In-depth understanding: A computer model of integrated processing for narrative comprehension*. Artificial Intelligence Series. MIT Press.
- Lehnert, W.G. (1981). Plot units and narrative summarization. *Cognitive Science*, 4:293-331.
- Löwe, B. (2010). Comparing formal frameworks of narrative structures. In M. Finlayson (Ed), *Computational Models of Narrative. Papers from the 2010 AAAI Fall Symposium*, (pp. 45-46). Volume FS-10-04 of AAAI Technical Reports.
- Löwe, B. (to appear). Methodological issues in comparing formal frameworks for narratives. In P. Allo & G. Primiero (Eds), *3rd Workshop on the Philosophy of Information*. Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.
- Marcu, D., Romera, M. & Amorrortu, E.A. (1999). Experiments in constructing a corpus of discourse trees: Problems, annotation choices, issues. In: *Workshop on Levels of Representation in Discourse*, (pp. 71-78).
- Marcus, M.P., Santorini, B. & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:302-330.
- Passonneau, R., Habash, N. & Ramnow, O. (2006). Inter-annotator agreement on a multilingual semantic annotation task. In: *Proceedings LREC-2006*.
- Propp, V. (1928). *Morfologiya skazki*. Leningrad: Akademija.
- Rumelhart, D.E. (1980). On evaluating story grammars. *Cognitive Science*, 4:313-316.
- Schank, R.C. (1982). *Dynamic memory: A theory of reminding and learning in computers and people*. Cambridge University Press.
- Turner, S. (1994). *The creative process. A computer model of storytelling*. Lawrence Erlbaum Associates.

## COMPUTERS AND PROCRASTINATION

*“I’ll just check my Facebook quick a Second...”*

NICK BREEMS

*Dordt College*

*Sioux Center, United States*

*and*

*University of Salford*

*Salford, United Kingdom*

**Abstract.** There seems to be something about computer technology that tempts us towards procrastination. This paper uses a philosophical toolkit to investigate why this might be, and how we can address the problem. We employ a framework for understanding the human use of computers developed by Andrew Basden. Basden's work is based on the thought of 20th century Dutch philosopher Herman Dooyeweerd, who makes the strong claim that reality is meaningful in a wide variety of mutually irreducible aspects. The non-reductionist approach of Dooyeweerd's philosophy allows Basden's framework to take everyday life seriously. Thus one of the strengths of a philosophical approach based on Dooyeweerd's thought is its ability to highlight important aspects of a problem that may be understudied. In this paper, the framework is used to perform an analysis of a particular example of computer-based procrastination, and potential avenues for investigation are highlighted that weren't immediately apparent when thinking about the problem generically. Thus we demonstrate that the use of a comprehensive framework for understanding the human use of computers and information systems from an everyday perspective shows some promise of providing insight into complex and challenging problems that arise in our information technology saturated culture.

### 1. Introduction

There seems to be something about computer technology and internet connectivity that distracts us, that tempts us towards procrastination. This is borne out by personal experience, by anecdotal evidence (Breems, 2009), and by research (Lavoie and Pychyl, 2001; Thatcher, Wretchko, and Fisher, 2008). For a tool widely believed to enhance our productivity, this is remarkable.

This naturally leads us to two questions:

1. Why is this?

2. How can we address this problem? What changes can we make in the way we design and implement computer systems or in the way we approach and use such technology that would reduce these distracting tendencies?

Research in the philosophy of computers and information systems can help us understand the use of computers as it plays out in everyday human living. This paper employs a framework for understanding the human use of computers developed by Andrew Basden (2008) in his book *Philosophical Frameworks for Understanding Information Systems*. We use this framework to analyze computer-induced procrastination, and demonstrate that philosophical tools can bring fresh insight to vexing problems.

## 2. Basden's Framework

In Chapter 4 of his book, Basden proposed a framework for understanding Human Use of Computers (the HUC framework), based work of 20th century Dutch philosopher Herman Dooyeweerd (1984). Dooyeweerd's thought is deeply non-reductionist: He made the strong claim that reality is meaningful in a wide variety of mutually irreducible aspects. Dooyeweerd identified a suite of fifteen such modal aspects, and posited that each of these aspects operates under a different set of laws which enable meaningful functioning in that aspect. Based on these insights, the HUC framework analyzes any particular use of computer technology along two axes. Horizontally, all computer use exists as three simultaneous functionings, because we're interacting with three different types of entity:

**Human/Computer Interaction (HCI)** To use a computer, we must interact with the computer itself: both with the hardware and with the user interface portions of the software.

**Engaging with Represented Content (ERC)** Computer programs represent content we engage with that is meaningful to us. For example, when we use an email program, it is not the internal voltages inside the CPU or the glowing of pixels on the screen that have direct meaning in our lives, but rather the content of the email messages and the information that they carry.

**Human Living with Computers (HLC)** The use of the computer plays out in our everyday lives; its effects escape the "box" that is the computer and affect things "out here" in our lived reality.

Vertically, he analyzes each of these functionings among each of Dooyeweerd's modal aspects:

**Quantitative** of discrete amount

**Spatial** of continuous extension

**Kinematic** of flowing movement

**Physical** of energy and mass

**Biotic/Organic** of life functions and integrity of organism

**Sensitive/psychic** of sense, feeling, and emotion

**Analytical** of distinction, conceptualizing, and inferring

**Formative** of formative power and shaping, in history, culture, creativity, achievement, and technology

**Lingual** of symbolic signification

**Social** of respect, social interaction, relationships, and institutions

**Economic** of frugality, skilled use of limited resources

**Aesthetic** of beauty, harmony, surprise, and fun

**Juridical** of “what is due”, rights, responsibilities

**Ethical** of self-giving love, generosity, care

**Pistic** of faith, commitment, trust, and vision

The non-reductionist approach of Dooyeweerd’s philosophy allows the framework to take everyday life seriously. That is, in our everyday experience of reality, we do not intuitively experience everything as mathematical, physical, or logical, but rather as diversely meaningful. The laws for the earlier aspects are largely descriptive; that is, we cannot disobey these laws (e.g. the law of gravity). The later laws, on the other hand, are prescriptive, and thus normative. They tell us how we *ought* to function, but do not force us to do so. For example, in the economic aspect, the law/norm of frugality tells us that we ought to use our time wisely. It allows us to make predictions about what kinds of consequences we can expect from obeying or not obeying that norm, but the choice to follow the norm or not is ours to make.

### 3. Use of the framework to analyze procrastination

One of the strengths of a philosophical approach such as Basden’s framework is its ability to highlight important aspects of a problem that may be understudied. In this paper, the framework is used to perform an analysis of a particular example of computer-based procrastination, playing an online dice game instead of writing a paper. Potential avenues for investigation are highlighted that weren’t immediately apparent when thinking about the problem generically:

- All of the dysfunction occurs in the HLC (Human Living with Computers) category, while most of the benefits of procrastinating (usually psychic and aesthetic) occur in the ERC (Engaging with Represented Content) functioning. Because ERC is a category that is much more within the control of a software designer, this points to the hope that design alternatives could help in addressing the problem.
- The proximity of the procrastinatory activity to the legitimate activity, both spatially and kinesthetically, eases the transition from real work to work avoidance. Although designing a computer to put physical distance between, for example, the use of a word processor and playing a game seems infeasible, there are potential designs which would increase the psychological distance from one activity to the other.
- The HLC functioning in the Pistic aspect indicates that procrastination is a failure of commitment: We are insufficiently committed to the course of action we are committed to, resulting in a break of faith with other people in our lives, our selves, and ultimately, with our religious convictions. A similar theme is suggested by Pychyl (2008).

Performing an analysis such as this, and evaluating the insight that results, is a preliminary way of testing the utility of the HUC framework itself. Thus we demonstrate that the use of a comprehensive framework for understanding the human use of computers and information systems from an everyday perspective shows some promise

of providing insight into complex and challenging problems that arise in our information technology saturated culture.

## References

- Basden, A. (2008). *Philosophical Frameworks for Understanding Information Systems*. Hershey, PA: IGI Publishing.
- Breems, N. S. (2009, September 8). Nick Breems is doing a short research project [web log post]. Retrieved from <http://www.facebook.com>
- Dooyeweerd, H. (1984) *A New Critique of Theoretical Thought* (Vols. 1-4). Jordan Station, Ontario, Canada: Paideia Press. (Original work published 1953-1958).
- Lavoie, J. A. A., & Pychyl, T. A. (2001). Cyberslacking and the procrastination superhighway: A web-based survey of online procrastination, attitudes, and emotion. *Social Science Computer Review* 19, (4), 431-444.
- Pychyl, T. A. (2008, April 7). Existentialism and procrastination: Bad faith. [Web log post]. Retrieved from <http://www.psychologytoday.com/node/372>
- Thatcher, A., Wretchko, G., & Fisher, J. (2008). Problematic internet use among information technology workers in South Africa. *CyberPsychology & Behavior* 11 (6), 785-787.

## COMBINATORY LOGIC WITH FUNCTIONAL TYPES IS A GENERAL FORMALISM FOR COMPUTING COGNITIVE AND SEMANTIC REPRESENTATIONS

JEAN-PIERRE DESCLÉS

*Laboratory LaLIC, University of Paris-Sorbonne  
Maison de la Recherche, 28 rue Serpente, 75006, Paris, France*

HEE-JIN RO

*Laboratory LaLIC, University of Paris-Sorbonne  
Maison de la Recherche, 28 rue Serpente, 75006, Paris, France*

AND

BRAHIM DJIOUA

*Laboratory LaLIC, University of Paris-Sorbonne  
Maison de la Recherche, 28 rue Serpente, 75006, Paris, France*

**Abstract.** We show how it is possible to use explicitly Combinatory Logic (a logic of operators and composition of operators) to define aspectual operators and temporal relations in natural languages from basic primitives in the domain of the temporality.

### 1. Combinatory Logic

Combinatory Logics with functional types (CL) is a formalism used for studying the foundations of Computer Sciences (semantics of Programming Languages) and for defining functional programming Languages (as HASKELL) built from this logical model. CL is a logic of operators and composition of operators. CL has been developed principally by Curry and Feys (1958), and then it has been used in linguistics by Shaumyan (1987) and by Desclés (1990).

In computer science, an applicative program is viewed as a combination of elementary programs, the program being built up with the help of a complex combinator, this latter being the result of an applicative combination of elementary combinators. The same idea can be used in other fields: logic and philosophy (logical analysis of paradoxes and some philosophical concepts), nanostructures synthesis and molecular combinatory computing (MacLennan, 2003), cognitive representations where a symbolic representation is an applicative organization of semantic primitives... Linguistic units are viewed as operators and operands of different functional types.

CL allows, on the one hand, to articulate, inside of a same computational architecture, different representation levels during a process of change of levels and, on the other hand, to give, by means of a formal calculus, a synthesis of a lexical (or grammatical) operator from its meaning.

## 2. Semantic Analysis of Aspecto-Temporal Operators

We present a semantic analysis of some aspectual and temporal operators. Grammatical units (aspects, tenses, moods ...) are operators whose meanings are analysed with elementary semantic operators combined together with a combinator. An aspectual operator ‘ASP<sub>I</sub>’ is applied onto a predicative relation ‘Λ’ (as “*Peter to enter the-room*” or “*Peter to be inside the room*”) where ‘I’ is a topological interval of contiguous and ordered instants, this interval specifying the temporal area of realization of ‘Λ’. There are three basic aspectual operators STATE<sub>O</sub>, EVENT<sub>F</sub> and PROC<sub>J</sub>. If an aspectualized predicative relation ‘ASP<sub>I</sub>(Λ)’ is viewed as a state ‘STATE<sub>O</sub>(Λ)’, then the interval ‘O’ is open and ‘Λ’ is true at every instant of ‘O’ (example (1) *Peter is inside the room* is a descriptive state). If ‘ASP<sub>I</sub>(Λ)’ is an event ‘EVENT<sub>F</sub>(Λ)’ ((2) *Peter entered the room*), the interval ‘F’ is closed and ‘Λ’ is always true at the final bound of ‘F’ (end of the complete event). If ‘ASP<sub>I</sub>(Λ)’ is a process ‘PROC<sub>J</sub>(Λ)’ ((3) *Peter is entering inside the room*), the interval ‘J’ is closed at the left bound of ‘J’ (beginning of the process) and open at the right bound of ‘J’ to mean that the process is uncomplete.

For speaking, the speaker must locate ‘ASP<sub>I</sub>(Λ)’ inside the temporal referential framework organized by himself; his speech act is an uncomplete process expressed by “I-AM-SAYING (...)” = “PROC<sub>J<sup>0</sup></sub>(I-SAY (...))”, where ‘J<sup>0</sup>’ is the interval of speaking, with its right open bound (the process of speaking is fundamentally uncomplete). The temporal intervals ‘O’, ‘F’ and ‘J’ can be related to the interval ‘J<sup>0</sup>’. For the examples (1), (2) and (3), we obtain the respective temporal relations between right bounds of different intervals:

$$[\delta(O) = \delta(J^0)] \quad (1')$$

$$[\delta(F) < \delta(J^0)] \quad (2')$$

$$[\delta(J) = \delta(J^0)] \quad (3')$$

where ‘δ’ and ‘γ’ are respective operators that selects the right and left bounds of an interval.

The combinators are used to express how the aspectual operators and temporal relations are combined together and synthesized into an unique grammatical operator expressed by a morphological operator. CL gives tools to analyze complex units into a combination of more elementary units. The computing of synthesis processes in a top-down strategy (or the analytic decomposition in a bottom-up strategy) of numerous aspectual and temporal operators has been realized with HASKELL. By the same way, the automatic analysis of some lexical predicates into a scheme where are combined semantic primitives in an applicative expression has been realized. We have not the place to show all steps of deductions for different aspectual operators which highlight the notions about process, event, state and related notions. With the adjunct of semantic representation of the lexical predicates, it becomes possible to give the formal deduction from a given sentence to another (Desclés, 2005; Desclés and Ro, 2011):

*John took the Mary's pen → Mary doesn't have the pen anymore*

When a speaker of English understands the first sentence, it is able to infer automatically the second sentence. This inference becomes possible with a grammatical knowledge (meaning of tenses) and a representation of the meaning of lexical predicate to take. Our research program shows how a machine can simulate this kind of inference realized by humans. For more details, to see (Desclés, 1990; 2005) and (Desclés & Ro, 2011a; 2011b).

## References

- Curry H.B. & Feys R. (1958). *Combinatory logic. Vol. I*. Studies in logic and the foundations of mathematics, North-Holland Publishing Co., Amsterdam
- Desclés J.-P. (1990). State, event, process, and topology. In: *General Linguistics* (pp.159-200), vol.29, n°3, Pennsylvania State University Press, University Park and London.
- Desclés J.-P. (2005). Reasoning and Aspectual-Temporal Calculus. In: Vanderveken D. (Eds), *Logic, Thought and Action*, Springer (pp. 217-244).
- Desclés J.-P. & Ro H.-J. (2011a). Aspecto-Temporal Representation for Discourse Analysis an Example of Formal Computation, *The 24th Florida Artificial Intelligence Research Society Conference*.
- Desclés J.-P. & Ro H.-J. (2011b). Operateurs aspecto-temporels et Logique Combinatoire. To appear in *Mathématiques et Sciences Humaines*.
- Hindley J.R. & Seldin J.P. (1986). *Introduction to Combinators and Lambda-Calculus*. Cambridge Univ. Press.
- MacLennan, B. J. (2003). *Combinatory Logic for Autonomous Molecular Computation*, [www.cs.utk.edu/~mclennan](http://www.cs.utk.edu/~mclennan)
- Shaumyan S.K. (1987). *A Semiotic Theory of Natural Languages*. Bloomington: Indiana University Press

## THE PAST, PRESENT, AND FUTURE ENCOUNTERS BETWEEN COMPUTATION AND THE HUMANITIES

STEFANO FRANCHI  
*Department of Hispanic Studies*  
*Texas A&M University*  
*stefano@tamu.edu*

**Abstract.** The paper addresses the conference theme from the broader perspective of the historical interactions between the Humanities and computational disciplines (or, more generally, the “sciences of the artificial”). These encounters have followed a similar although symmetrically opposite “takeover” paradigm. However, there is an alternative meeting mode, pioneered by the interactions between studio and performance arts and digital technology. A brief discussion of the microsound approach to musical composition shows that these alternative encounters have been characterized by a willingness on both parts to let their basic issues, techniques, and concepts be redefined by the partner disciplines. I argue that this modality could (and perhaps should) be extended to other Humanities disciplines, including philosophy.

### 1. Takeovers

The two best-known encounters between computational technologies and traditional Humanists pursuits are represented by the Artificial Intelligence/Cognitive science movement and the roughly contemporary Digital Humanities approach (although the label became popular only recently). Classic Artificial Intelligence saw itself as “anti-philosophy” (Dupuy, 2000; Agre, 2005; Franchi, 2006): it was the discipline that could take over philosophy’s traditional questions about rationality, the mind/body problem, creative thinking, perception, etcetera, and solve with the help of a set of radically new synthetic, experimental-based techniques. The true meaning of the “computational turn in philosophy” lies in its methodology, which allowed it to associate engineering techniques with age-old philosophical questions. This “imperialist” tendency of cognitive science (Dupuy, 2000) was present from the very beginning, even before the formalization of the field into well-defined theoretical approaches (McCulloch (1989[1948]); Simon, 1994).

The Digital Humanities represent the reverse modality of the encounter just described. The most common approach (Kirschenbau, 2010) uses tools, techniques, and algorithms developed by computer scientists to address traditional questions about the meaning of texts, their

accessibility and interpretation, and so on. Other approaches turn technology into the scholar's preferred object of study (Svensson, 2010). The recent approach pioneered by the "Philosophy of Information" (Floridi, 2011) follows this pattern. Its focus on the much broader category of "information" substantially increases the scope of its inquiries, while firmly keeping it within philosophy's standard reflective mode.

The common feature of these two classic encounters between the Humanities and computational theory and technology is their one-sidedness. In either case, one of the two partners took over some relevant aspects from the other participant and fit it within its own field of inquiry (mostly questions, in AI's case; mostly tools, for the Digital Humanities). The appropriation, however, did not alter the theoretical features of either camp. For instance, AI and Cognitive Science researchers maintained that philosophy pre-scientific methodology had only produced mere speculation that made those problems unsolvable. Therefore, philosophy's accumulated wealth of reflection about the mind, rationality, perception, memory, emotions, and so forth could not be used by the computational approach. In McCulloch's famous phrase, the "den of the metaphysician is strewn with the bones of researchers past." In the Digital Humanities' case, the takeover happens at the level of tools. In most cases, however, this appropriation does not become an opportunity for a critical reflection on the role of the canon on liberal education, or for a reappraisal of the role of the text and the social, political, and moral roles it plays in society at large.

## 2. Digital practice

Meetings between artists and computational technology show the possibility of a different paradigm. In many cases, making music, painting, producing installations, and writing with a computer changes the concepts artists work with, and, at the same time, forces computer sciences to change theirs as well. There are many examples in the rich history of "digital art," broadly understood (OuLiPo, 1973; ALAMO, No year; Schaeffer, 1952). I will illustrate their general features with reference to a more recent project: the "microsound" approach to musical composition (Roads, 2004).

"Microsounds" are sonic objects whose timescale lies between that of *notes*—the smallest traditional music objects, whose duration is measured in seconds or fractions thereof—and *samples*—the smallest bit, measured in microseconds ( $10^{-6}$ ). The manipulation of microsounds broadens substantially the composer's palette, but it is

impossible without the help of technological devices of various kinds, from granular synthesis software to high-level mixing interfaces. Composers wishing to “sculpt” sounds at the microlevel face a double challenge that translates into a mutual collaboration between compositional and algorithmic techniques. On the one hand, they need to broaden the syntax and grammar of music's language to allow the manipulation and aesthetic assessment of previously unheard of objects (Vaggione, 2001). On the other hand, they need computer scientists and mathematicians to develop alternative analytic and synthetic models of sound (in addition to Fourier-transforms and similar methods) capable of capturing the features of sonic events lasting only a few milliseconds (Vaggione, 1996).

This example of artistic production points to a pattern of cooperation between work in computational and non-computational disciplines that is deeply at odds with the AI/CogSci and DigHum patterns discussed above. Instead of a takeover, the artistic model produces a true encounter that changes both partners' technical and theoretical apparatus.

### 3. Posthuman encounters?

Could the encounter model practiced by artists be generalized to the Humanities? We can see how this could be the case by considering a twofold question. On the one hand: are Humanities' traditional inquiries about human nature and human cultural production still relevant in a landscape in which some of the communicating agents may not be human, partially or entirely? Can they go on in the same way? And *vice versa*: are science and technology fully aware that the new digital artifacts they are shepherding into the world may change its landscape and transform worldly action at the pragmatic as well as at the theoretical level? Or are they still relying upon a pre-digital universe in which technological artifacts were always to be used as mere tools deployed by humans, an assumption that seems increasingly questionable?

I think a particularly fruitful approach toward this question is provided by the kind of critical thought that has been developed—mostly, but certainly not exclusively—in Continental Europe over the last two or three decades. These theoretical efforts have based their explorations upon anti-humanist and/or post-humanist perspectives. They provide, therefore, a fruitful starting point for the investigation and interaction with instruments, tools, and techniques that question the very notion of the human. For instance, Lacanian and post-Lacanian psychoanalysis has articulated a view of the human that deploys cybernetic concepts to explain high level cognitive functions (Franchi, 2011; Chiesa, 2007);

the work on biopolitics currently developed by largely Italian philosophers attempts to articulate a conception of human life that is continuous with animal and non-organic life (Agamben, 2003; Esposito, 2008; Tarizzo, 2010). At the same time, the disciplines of science and technology studies in their contemporary North American, French, and German developments have provided penetrating analyses of the bidirectional relationships between scientific theories and technological artifacts, on the one hand, and philosophical and cultural productions on the other (Ihde, 2002; Hayles, 1999; Latour and Woolgar, 1986; Biagioli, 1999).

This suggestion does not pretend to exhaust the theoretical options we have at our disposal when reflecting upon the computational turn. My contention, however, is that artistic practices in all forms of “digital art” can serve as an inspiration to all of the Humanities disciplines. We can follow their path toward a new mode of digital encounter that does not fall into the well-worn path of hostile takeovers by either partner.

### References

- Agamben, G. (2003). *The Open. Man and Animal*. Stanford, Calif.: Stanford University Press.
- Agre, Ph. E. (2005). The Soul Gained and Lost: Artificial Intelligence as Philosophical Project. In: S. Franchi and G. Güzeldere (Eds.), *Mechanical Bodies, Computational Minds* (pp.153-174). Cambridge: MIT Press.
- ALAMO (*Atelier de Littérature Assistée par la Mathématique et les Ordinateurs*). Url: <http://alamo.mshparisnord.org/index.html>
- Biagioli, M. (Ed.) (1999). *The Science Studies Reader*. New York: Routledge.
- Chiesa, L. (2007). *Subjectivity and Otherness. A Philosophical Reading of Lacan*. Cambridge: MIT Press.
- Dupuy, J.-P. (2000). *The Mechanization of the Mind: On the Origins of Cognitive Science*. Princeton: Princeton University Press.
- Esposito, R. (2008). *Bios: Biopolitics and Philosophy*. Minneapolis: University of Minnesota Press.
- Floridi, L. (2011). *The Philosophy of Information*. Oxford: Oxford University Press.
- Franchi, S. (2006). “Herbert Simon, Anti-Philosopher.” In: L. Magnani (Ed.) *Computing and Philosophy* (pp. 27-40). Pavia: Associated International Academic Publishers.
- (2011). Jammed Machines and Contingently Fit Animals: Psychoanalysis’s Biological Paradox, *French Literature Series*, 38, in press.
- Hayles, N. K. (1999). *How We Became Posthuman: Virtual Bodies in Cyberspace*. Chicago: University of Chicago Press.
- Ihde, D. (2002). *Bodies in Technology*. Minneapolis: University of Minnesota Press.
- Kirschenbau, M. G. (2010). What Is Digital Humanities and What’s It Doing in English Departments? *ADE Bulletin*, 150, 1–7.
- Latour, B. and Woolgar, S. (1986). *Laboratory Life: the Construction of Scientific Facts*. Princeton: Princeton University Press.

- McCulloch, W. S. (1989[1948]). Through the Den of the Metaphysician. In: *Embodiments of Mind* (142-156). Cambridge: MIT Press.
- OuLiPo (1973). *La littérature potentielle*. Paris: Gallimard.
- Roads, C. (2004). *Microsound*. Cambridge: MIT Press.
- Schaeffer, P. (1952). *À la recherche d'une musique concrète*. Seuil.
- Simon, H. (1994). Literary Criticism: a Cognitive Approach. In: S. Franchi and G. Güzeldere (Eds.), *Bridging the Gap* (pp. 1–26). *Stanford Humanities Review*, 4(1), Special Supplement.
- Svensson, P. (2010). The Landscape of Digital Humanities. *Digital Humanities Quarterly*, 4(1).
- Tarizzo, D. (2010). *La vita, un'invenzione recente*. Bari: Laterza.
- Vaggione, H. (1996). Articulating Microtime. *Computer Music Journal*, 20(2), 33–38.
- (2001). Some Ontological Remarks about Music Composition Processes. *Computer Music Journal*,

## REFLECTIONS ON NEUROCOMPUTATIONAL RELIABILISM

MARCELLO GUARINI

*Department of Philosophy, University of Windsor  
401 Sunset, Windsor, ON, Canada N9B 394*

AND

Joshua Chauvin

and

Julie Gorman

*Students, Department of Philosophy, University of Windsor  
401 Sunset, Windsor, ON, Canada N9B 394*

### 1. Introduction

Reliabilism is a theory of knowledge that has traditionally focused on propositional knowledge. Paul Churchland has advocated for a reconceptualization of reliabilism to “liberate it” from propositional attitudes (such as accepting that  $p$ , believing that  $p$ , knowing that  $p$ , and the like). In the process, he (a) outlines an alternative for the notion of truth (which he calls “representational success”), (b) offers a non-standard account of theory, and (c) invokes the preceding ideas to provide an account of representation and knowledge that emphasizes our skill or capacity for navigating the world. Crucially, he defines reliabilism (and knowledge) in terms of representational success. This paper discusses these ideas and raises some concerns. Since Churchland takes a neurocomputational approach, we discuss our training of neural networks to classify images of faces. We use this work to suggest that the kind of reliability at work in some knowledge claims is not usefully understood in terms of the aforementioned notion of representational success.

### 2. Traditional Reliabilism: Truth and Propositional Attitudes

Claims to propositional knowledge have the form, *S knows that p*, where  $p$  is a proposition. For the reliabilist, among the necessary conditions for some agent or subject  $S$  to know  $p$  are that (a)  $p$  is true, (b)  $S$  believes  $p$ , and (c)  $p$  is the outcome of a reliable process or method. According to Alvin Goldman (1986, 1992, 1999, 2002) reliability is required for both epistemic justification and knowledge. This reliability is a ratio: the number of true beliefs delivered by a process or method divided by the number

of true and false beliefs delivered by the same process or method. As we will concern ourselves primarily with the reliability requirement in this paper, we shall not engage the issue of what might constitute sufficient conditions for either knowledge or justification.

### **3. Neuro Reliabilism: Representational Success and Similarity Spaces**

Paul Churchland (2007) attempts to take a reliabilist approach to epistemology, divorce it from propositional attitudes, and explain how we can have non-propositional knowledge. Churchland begins by enumerating many instances of know-how. The examples include the capacity or skill knowledge possessed both by humans and non-humans. He argues that much of what we call knowledge has little or nothing to do with the fixing of propositional attitudes. He recognizes the importance of truth in classical approaches to reliabilism, but he resists talking of truth since (a) it attaches to propositional attitudes, and (b) much of our knowledge is not about fixing propositional attitudes. In place of truth, Churchland formulates a notion of representational success that is compatible with analyses of neural networks. To keep things simple, consider a three layer feed forward neural network. After training, each different pattern of activation across the hidden units is a different point in that space. We can then measure the distance between points (which Churchland often refers to as similarity relations). Churchland treats (somewhat metaphorically) similarity spaces as maps that guide our interactions with the world. Just as a map is representationally successful when the distance relations on the map preserve distance relations in the world, conceptual spaces understood as similarity spaces are representationally successful when they preserve similarity or distance relations between points in state space and the world.

### **4. How Representational Success and Reliability can Come Apart**

We will present the results of two neural networks (N1 and N2) trained to classify images of faces as either male or female. N1 was trained on the set of images A; it was tested on images it had not previously seen, set B. N2 was trained on B; it was tested on A. Both networks achieved equal levels of success on the images. In spite of the preceding, we will show that N1 and N2 set up different similarity spaces. This is a problem for Churchland's position since he defines reliability in terms of representational success, and this latter notion is defined in terms of structure preserving mapping between points in similarity space and features of the world. It seems quite natural to say that N1 and N2 are equally reliable, but because they set up different similarity spaces, we will argue that it is not clear how they could be equally representationally successful, given the work Churchland expects representational success to do.

There is a difference between (a) being reliable and (b) explaining the source of that reliability. We will show that we can understand what it is for a system (a face classifying neural network) to be reliable independent of understanding the source of that reliability. Churchland uses the notion of representational success (or preservation of distance relations) both to define reliability and to understand its source (i.e. to do both (a) and (b)). This is a source of potential problems for his position.

## 5. Conclusion

In spite of the problems, we recognize there are some attractions to the sort of position Churchland is putting forward. While we do not think it has the range of applicability Churchland suggests, we do not take ourselves to have argued that representational success is a useless notion. We will close with some constraints that need to be satisfied for the notion to be a useful one.

## Acknowledgements

We thank the Shared Hierarchical Academic Research Computing Network (SHARCNet) for financial support.

## References

- Churchland, P.M. (2007). *Neurophilosophy at Work*. Cambridge, UK: Cambridge University Press.
- Goldman, A. (1986). *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.
- Goldman, A. (1999). *Knowledge in a Social World*. Oxford: Oxford University Press.
- Goldman, A. (1992). *Liaisons: Philosophy Meets the Cognitive and Social Sciences*. Cambridge, MA: MIT Press.
- Goldman, A. (2002). *Pathways to Knowledge, Private and Public*. Oxford: Oxford University Press.

## STATES OF AFFAIRS AND INFORMATION OBJECTS

STEVE T. MCKINLAY  
*Charles Sturt University*

*Wellington Institute of Technology  
School of Information Technology  
Private Bag 39803, Petone, Wellington, NEW ZEALAND  
e-mail: steve.mckinlay@weltec.ac.nz*

**Abstract.** This paper compares two recently detailed metaphysical accounts of reality. On the one hand we have Luciano Floridi's "information realism" and, on the other David Armstrong's view that the general structure of reality can be described as "states of affairs".

Floridi postulates the information object as the entity central to information ethics and his *informational realism*. In developing the concept he draws heavily upon object oriented (OO) programming theory. Informational objects are reckoned by Floridi to be, in a sense, ontologically primitive and as such naturally occurring mind independent structures dynamically interacting with one another. Floridi employs OO like terminology such as "properties" and "relations" in order to clarify his concept of the informational entity.

Armstrong on the other hand postulates that the world, all that there is, is a *world of states of affairs*. A state of affairs according to Armstrong consists of a particular, which has a property or alternatively a relation which holds between two or more particulars. Each state of affairs as well as constituent higher or lower order states of affairs is a contingent existent. Furthermore the properties and relations attached to states of affairs are universals.

These two theories, whilst exhibiting marked resemblances also reveal fundamental philosophical differences yet both attempt to present a unified metaphysical schema, an ontology. Of great interest is the fact that here we have two strong competing theories. The situation begs critical comparison. Such a comparison is the primary aim of this paper.

The idea of the *Information Object* as being somehow ontologically fundamental has gained traction recently not only in computer programming circles but also philosophically. We could attribute this newfound popularity, particularly with regard to philosophical interpretations, with the fact that we live in the so called *information age*. We, at least in the developed world, view the world through information-coloured spectacles these days. Adding some substance to this claim is the fact that our information systems are designed and developed using fashionable object oriented (OO) methodologies. Information modeling is now the accepted process by which facts or propositions, the sentences that demarcate the various states of affairs and "things" of

which the modeller is interested, are defined via “object class” structures. Such structures in turn represent various properties, behavior and *relata*.

The information object in this sense is an intuitively fitting and elegant way of representing the problems we attempt to solve via computational means. OO design and development is “instrumentally reliable” – it works. The majority of modern implemented information technologies across the entire gamut of industries and applications typically employ object oriented approaches. The focus has shifted from procedural algorithmic processing to an object driven methodology and as such states of affairs and “things” are abstractly modelled as self-contained (encapsulated) object structures, responsible for their own identity, relations, properties, states and behavioural rules. It’s perhaps not surprising then that we might ponder; could the universe be interpreted and/or represented in such a way?

From a wider perspective what is often termed the *computational turn* has given rise to the informational object concept central to and emerging as fundamental in an informational ontology developed primarily by Luciano Floridi (2002, 2004, 2008). The concept is important for Floridi since the information object plays a role central to his Information Ethics (IE) and Informational Realism (IR). But more than this, the idea of the “information entity” seems to offer new ways of understanding epistemology, semantics, scientific explanation, and ethics. Floridi has developed a detailed picture of the information object (or entity as he sometimes calls it) employing Object Oriented programming and design methods and theories to clarify the concept.

Whilst Luciano Floridi’s notion of the information object is somewhat analogous to the OO conception of an object in a recent paper I argued for a variety of reasons that information objects, certainly within the context of Floridi’s informational realism, don’t seem to be much like OO objects, certainly not the kind employed in an OO class model or an OO program. Arguably the most significant difference is that OO objects act unequivocally as referents to *facts*, as Wittgenstein (1961) would have put it, or what Armstrong (1998) calls *states of affairs*. I think there is certainly a similarity between OO objects and Floridi’s conception of the information object but I suspect the similarity is more harmful to the idea of the information object holding any independent ontological status or existing independently as a particular category. The similarity is that both object concepts are largely conceptual by nature. Yet Floridi seems to want to confer a stronger ontological status to the information entity. Problems arise if the information object is indeed conceptual. Following Lauden (1977, p48) such entities can have no existence independent of the theories within which they are postulated.

Nevertheless the concept of an information entity is certainly a convenient and relatively intuitive way of bundling up constituent properties and relations belonging to the *particular* in question. Those properties and relations are in fact what philosophy sometimes calls universals and it is each particular (distinct information objects) that instantiate those universals. The universals themselves are the constituents of information objects shared across many objects. There are some that deny the existence of universals (nominalism) and we shall consider this in the full paper.

Armstrong (1998, p95) questions the need to recognise an independent category of particulars. He argues that whilst properties and relations can be known “the bearer of properties and relations, it is alleged, cannot be known. Why then postulate a bearer?” The postulation of bearers, Armstrong argues, appears to lack ontological and epistemic

economy (*ibid*). This raises the question, is the Floridian information object the same kind of thing Armstrong terms a *bearer*?

From the OO perspective a particular information object (or class, although the two concepts differ slightly and this will be explained) is admittedly *representative* of a fact, state of affairs or physical object, this renders the OO object second order to the actual fact or state of affairs. Furthermore I take it, it is meant to be information objects all the way down. But we already see this isn't the case. Information objects are essentially bundles of properties and relations, whilst no information object can be strictly identical with another, the properties and relations can and are identical across multiple instantiations of similar objects. Whilst they do not exist outside their instantiations it would seem properties and relations hold a more fundamental ontological position than the information entity.

Thus to uphold the ontological reality of "information objects" or in Armstrong's case "states of affairs" seems to entail the admission of properties and relations yet there would certainly be some philosophers who would deny that the reverse holds. There seems to be little controversy in the admission of properties and relations since a denial results in the denier having to come up with an alternate theory of classes. It is individual objects or states of affairs that exhibit more or less identical properties and relations that we bundle into classes.

This paper compares Armstrong's descriptions of properties and relations with those affiliated to Floridi's information object concept. Further we will consider how similar (or different) the information object concept is to the Armstrong's conception the state of affairs.

## References

- Armstrong D.M. (1989). *Universals: An Opinionated Introduction*. Westview Press. (Focus Series)
- Armstrong, D. M.. (1998). *A World of States of Affairs*. Cambridge: Cambridge University Press.
- Floridi, L. (2002). On the Intrinsic Value of Information Objects and the Infosphere. *Ethics and Information Technology*, 3(4), 287-304.
- Floridi, L.. (2004). Informational Realism. In G.M. Greco, *IEG Research Report* Oxford: Information Ethics Group.
- Floridi, L. (2008). A Defence of Informational Structural Realism. *Synthese*, 161(2), 219-253.
- Lauden, L.. (1977). *Progress and Its Problems: Towards a Theory of Scientific Growth*. California: University of California Press.
- Wittgenstein, L.. (1961). *Tractatus Logico-Philosophicus*. London and New York: Routledge.

## SCIENTIFIC EXPLANATION AND INFORMATION

STEVE T. MCKINLAY

*Charles Sturt University*

*Wellington Institute of Technology*

*School of Information Technology*

*Private Bag 39803, Petone, Wellington, NEW ZEALAND*

*e-mail: steve.mckinlay@weltec.ac.nz*

**Abstract.** *Scientific explanation* and more recently *information* have attracted considerable philosophical attention. Little consideration however has been given to making sense of the concept of information used within debates surrounding explanation. Some may deem there is no problem to be solved here. Yet we observe within the literature on scientific explanation strict examinations of profound philosophical concepts. Writers are at pains to explain causal, epistemic, ontological and nomological accounts of explanation all of which in some way rely upon and take for granted the role of information.

We like to think these days we have, at least the beginnings of, a coherent theory of information. This paper cherry picks a couple of interesting ideas within scientific explanation and attempts to reconcile the generally received view of information with those particular explanatory accounts. By the received view I mean the *General Definition of Information* mostly attributed to Luciano Floridi from around 2003 onwards. As a result of this investigation some profound questions arise; is an “ideal explanatory text” (see Railton, 1981) essentially an informational concept? Can we make sense of a relationship between causation and information? Just how are the concepts related and do we need a satisfactory account? And finally, is it possible to propose a purely information-centric theory of scientific explanation and if so, could it be a significant improvement on current theories of scientific explanation?

*Everything that exists makes a difference to the causal powers of something.*  
*David Armstrong, 1997, p. 41)*

### Introduction

Wesley Salmon in *Causality and Explanation* suggests that to most people, the fact that there is a close connection between causality and explanation comes as no surprise (1998, p. 3). And while distinctions can

certainly be made between the two concepts there are many convergences. Salmon argues, “In many cases to explain something is to state its cause.” (*ibid*). I happen to think a similar story can be told with regard to information and explanation. To have something explained is, at least from an ordinary language point of view, to be informed. There is a certain structure about scientific explanation, the various relationships between laws and theories, and information seems to be the flesh on these bones. It follows that the concept of information might benefit from an investigation into the connections or relations that exist between it, causal concepts and explanation and it is this particular can of worms that this paper intends to open.

### **Information, Causality and Explanation**

The body of philosophical literature on scientific explanation is substantial beginning<sup>16</sup> with the deductive-nomological (D-N) model (Hempel & Oppenheim, 1948., Hempel, 1965) wherein scientific explanations were considered deductive arguments<sup>17</sup>. Salmon (1971) followed with the statistical relevance (S-R) model in order to deal with explanations of low probability events not adequately dealt with by Hempel’s explanatory models. Later Railton (1978, 1981) proposed a deductive-nomological-probabilistic (D-N-P) model in a further attempt to explain events that happen by chance. More recently Wesley Salmon proposed a casual theory of explanation.

Salmon’s principal claim was that a scientific explanation is constituted by a state of affairs predominantly recognised as a pattern in the world where that pattern consists of at least one causal process. Causal processes Salmon argued (also Railton, 1981 and later Dowe, 2000) *necessarily* transmit information (1998, p.16). Salmon explains this as the ability of a causal process to transmit a mark. Causal processes are described by Salmon as being continuous (in a physically spatio-temporal way). This view contrasts with the popular view of causality being a “relation” between particular events (the *cause*, and the *effect*). Salmon’s theory is perhaps most eloquently clarified in his *At-At Theory*

---

<sup>16</sup> Although the roots of scientific explanation and understanding can of course be traced back well beyond Aristotle, recent philosophical history regarding scientific explanation is generally considered to begin with Hempel and Oppenheim’s ground breaking paper *Studies in the Logic of Explanation*.

<sup>17</sup> The degree of informativeness of a logically deductive schema is perhaps controversial, however given scientific explanation has moved on considerably from the Hempelian D-N approach we can safely leave this controversy to one side.

of *Causal Influence* (1977, reprinted in Salmon, 1998). The At-At theory Salmon claims not only resolves Zeno's arrow paradoxes but also proposes a foundation for a concept of propagation of causal influence. Information plays a significant yet largely unexplained role in virtually all of the models of explanation particularly Salmon's At-At causal theory.

The usual constraints prevent this paper from adequately summarising in full the development of scientific explanation from Hempel's D-N model through recent attempts at a unified model of explanation and so I intend to choose two particular junctures in the history of scientific explanation in the hope of casting some light upon the controversial three way axis between *information, causality* and *explanation*. As is often the case in philosophy the following investigation is most likely to end in more, yet hopefully new and interesting questions regarding the nature of information. Thus, my two starting points with their associated problems are as follows;

1. Peter Railton makes a distinction between what he terms the "ideal explanatory text" and "explanatory information" (1981, p. 240). Railton openly admits in his 1981 paper that whilst it is typical to speak of sentences or texts conveying information he knows of "no satisfactory account of this familiar and highly general notion" (1981, p. 240). Further he admits that neither does the notion of information defined by Wiener and Shannon appear to fit his explanatory theory. Given that Railton's work continues to influence attempts at theories of explanation, in particular Kitcher's (1989) unificationist account, an enquiry into Railton's "explanatory information" seems overdue.
2. Wesley Salmon's development of Reichenbach's "mark method" in his At-At Theory of Causal Influence makes thoughtful claims about information transmission as a result of causal processes. Salmon makes a clear distinction between causal processes and pseudo-processes, the latter he claims have no ability to transmit information. I will evaluate Salmon's claims with examples and examine how Salmon's concept of information transmission squares with our current views about information.

This investigation I think raises profound questions; is Railton's concept of the ideal explanatory text essentially an informational concept? On the other hand can we make sense of a relationship between causation and information? Just how are these concepts related and do we need a

satisfactory account? Finally, can we propose an informationally centred theory of scientific explanation? Rather than attempt to conclusively answer these questions in this paper, I hope to build an argument around the fact that the topic is one worthy of serious consideration.

## References

- Dowe, P.. (2000). *Physical Causation*. Cambridge: Cambridge University Press.
- Floridi, L.. (2003). From Data to Semantic Information. *Entropy*(5), 125-145.
- Hempel, C.. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.
- Hempel, C.. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, 15, 135-175.
- Kitcher, P.. (1989). Explanatory Unification and the Causal Structure of the World. In *Scientific Explanation* (410-505). Minneapolis: University of Minnesota Press.
- Railton, P.. (1981). Probability, Explanation, and Information. *Synthese*, 48(2), 233.
- Railton, P.. (1978). A Deductive-Nomological Model of Probabilistic Explanation.. *Philosophy of Science*, 45, 206-226.
- Salmon, W.. (1998). *Causality and Explanation*. Oxford: Oxford University Press.

## **BIOLOGICAL INSPIRED SINGLE-CHIP MASSIVELY PARALLEL SELF-HEALING, SELF-REGULATING, TERA-DEVICE COMPUTERS**

*Philosophical Implications of the Efforts for Solving Technological Show-Stoppers in the Path of the Next Computational Turn*

MICHAEL NICOLAIDIS  
*TIMA Laboratory*  
(CNRS, Grenoble INP, UJF)

**Abstract.** Biologically inspired computing usually addresses computing functionalities inspired from biological systems (genetic algorithms, neural networks, cellular automata, artificial life, ...). However, living organisms also resolve efficiently some other problems that have to be addressed in order to accomplish the next computational turn, achieving the robustness (reliability and power-dissipation) enabling making useful computations by means of ultimate CMOS (to be reached by the beginning of the next decade) and post-CMOS technologies. Thus, biologically inspired robust computing can be viewed as an emerging topic of biologically inspired computing. Complex organisms have the remarkable property of *self-healing*. Two fundamental features are on the basis of this ability. Organisms are constituted of large numbers of basic units (cells). Cells surrounding injured parts can substitute the dead cells and regenerate the damaged structures. Also, the cells themselves can recover from various damages, for instance by repairing their DNA. Furthermore, living organisms *regulate* their physiological parameters to the changing external conditions and their own needs (e.g. the regulation of insulin levels in response to sugar levels). As another remarkable property, the *autonomic nervous system* of higher animals controls important bodily functions (e.g. respiration, heart rate, and blood pressure) without conscious intervention. Building computers having similar properties and achieving the robustness they confer is an old dream of computer scientist. But so far, related researches did not lead to a practical self-healing, self-regulating, autonomic computing paradigm.

### **Ultimate CMOS and post-CMOS promises and challenges.**

We argue that today there are several converging factors which pave the way towards a new computing paradigms realizing this old dream. *These factors are three-fold. Two of them* are related with the technology scaling.

- Ultimate-CMOS and post-CMOS technologies promise integrating trillions devices in a single chip. Thus, single-chip *massively parallel architectures become mandatory* for utilizing the huge numbers of devices integrated in such chips.

- At the same time, aggressive technology scaling impacts dramatically process, voltage and temperature (PVT) variations; sensitivity to electromagnetic interferences (EMI) and to atmospheric radiation (neutrons, protons); and circuit aging; and also imposes stringent power dissipation constraints. The resulting high defect levels, heterogeneous behavior of identical processing nodes, circuit degradation over time, and extreme complexity, affect adversely fabrication yield and also prevent fabricating reliable chips in ultimate CMOS and post-CMOS technologies. These issues are the main show-stoppers in the path towards these technologies that pave the way for the next computational turn.

The **above two factors** plead for a self-healing massively parallel computing paradigm. But, this is not a trivial task. Copying with failures (a property also known as fault tolerance) induces high area and power penalties. The former will drastically reduce the available computing resources, while the later is incompatible with low power operation (one of the tightest constraints in ultimate CMOS). Furthermore, conventional fault-tolerant approaches (DMR, TMR etc) consider that failures affect a single component among several redundant ones. This assumption is no more valid in the extreme integration of ultimate CMOS, where transistors are so small that comprise a few atoms, neither under the even higher integration levels promised by post-CMOS. In these technologies we may face the following challenges:

- All processing nodes and routers in a massively parallel tera-device processor are affected by timing or transient faults,
- Hard faults may affect some parts of each node,
- Hard faults completely destroying a new node arrive every few days,
- Circuit degradation is continuous and requires continuous self-regulation of circuit parameters (clock-frequency, voltage levels, body bias), to maintain it operational.

### **Biologically-inspired enabling approaches**

*The Cells framework (On-Chip Self-healing Tera-Device Processors) discussed in this paper brings-in the third factor:* a drastically new system-design paradigm achieving high yield, and highly-reliable uninterrupted operation for highly defective on-chip massively parallel tera-device processors at low hardware cost. Power reduction and enhanced performance are also achieved through *self-regulation of circuit parameters* (voltage, clock frequency and body bias). Groundbreaking innovations were introduced at all levels of the *framework*, including its overall architecture, its particular components, and the way the cooperation of these components is architected to optimize the outcome. They enable continuous adaptation to circuit degradation, heterogeneity and changing application context, as well as detection and correct operation restoration for all failures induced by high defect densities, PVT variations, internal and external disturbances, and circuit degradation over time. It results in a holistic self-healing self-regulating approach allowing:

- Making usable tera-device technologies affected by: high defect densities, sever variability, increasing sensitivity to disturbances and accelerated aging.
- Implementing single-chip massively parallel self-healing tera-device computers delivering unprecedented computing power, which enable changing our computing paradigms and should have a profound impact on all computer application domains (including embedded systems, telecommunication networks, internet infrastructure and utilization, cloud computing, ...), as well as science and technology and the society as a whole.

In the *Cells*, *Self-Healing* is achieved by two means. Single-chip massively parallel processors resemble to living organisms in that they are constituted of large numbers of basic units (processor cores, routers and links). *Cells* takes advantage of this similarity. Like cells in living organisms, operational units replace unrecoverable faulty units to restore system functionality transparently to the ongoing application executions. Also, like cells in living organisms, processor cores, routers and links are able to recover from several kinds of failures, by using new innovations at circuit-level fault tolerance (Anghel and Nicolaidis, 2000), (Nicolaidis 2005), (Anghel and Nicolaidis, 2008), (Nicolaidis, 2011), (Yu, Nicolaidis, Anghel and Zergainoh, 2011) and self-regulation.

Furthermore, similarly to the non-deterministic, local and opportunistic manner in which cells in an organism achieve self-healing, and self-regulation, *Cells* uses new, non-deterministic routing, task allocation and scheduling algorithms, which make local decisions in opportunistic manner (Chaix, Avresky, Zergainoh and Nicolaidis, 2010 and 2011). They allow addressing the complexity problem of navigating in a complex and changing network (thousands of processors and routers, millions of possible communication paths, continuous circuit degradation, frequent occurrence of catastrophic node and router failures, and unpredictable router congestions). Conventional deterministic algorithms used in nowadays massively parallel multi-chip systems, which exhibit low defectivity and high circuit stability; use static routing tables containing pre-established routes, and static scheduling and allocation algorithms which consider: fixed clock frequencies; rarely failing links, router and processor nodes; and similar power-dissipation for all nodes. Such algorithms, used also in early proposals for designing massively parallel processor chips (Zajac, Collet and Napieralski, 2008), are ineffective in a highly defective and fast degrading hardware.

Together with the highly innovative circuit-level fault-tolerance, routing, and task allocation and scheduling; automatic monitoring, control, and *self-regulation* of circuit parameters ensure optimal operation: meeting performance requirements while minimizing power under circuit degradation and evolving application context.

It results in a computing paradigm that achieves robustness in a manner that resembles to biological systems in multiple aspects. This trend should be necessarily reinforced as post CMOS will enable ever higher integration complexities.

## References

- Anghel, L. & Nicolaidis, M. (2000), Cost Reduction and Evaluation of a Temporary Faults Detecting Technique, Proceedings Design Automation and Test in Europe Conference, March 2000, Paris (Best Paper Award)
- Anghel, L. & Nicolaidis, M. (2008), Cost Reduction and Evaluation of a Temporary Faults Detecting Technique”, chapter in the book “The Most Influential Papers of 10 Years DATE”, Lauwereins, Rudy; Madsen, Jan (Eds.), Springer, ISBN: 978-1-4020-6487-6, 2008.
- Chaix, F., Avresky, D., Zergainoh, N. E. & Nicolaidis, M. (2010), Fault-Tolerant Deadlock-Free Adaptive Routing for Any Set of Link and Node Failures in Multi-Cores Systems, In Proc. 9<sup>th</sup> IEEE International Symposium on Network Computing and Applications (NCA10), July 15-17 2010, Cambridge, MA
- Chaix, F., Avresky, D., Zergainoh, N. E. & Nicolaidis, M. (2011), A Fault-Tolerant Deadlock-Free Adaptive Routing for On Chip Interconnects, In Proc. Design Automation and Test in Europe Conference, March 14 – 18, 2011, Grenoble, France.
- Nicolaidis M., (2005), Design for Soft-Error Mitigation, IEEE Transactions on Materials and Device Reliability, Vol. 5, Issue 3, pp. 405-418, September 2005

- Nicolaidis, M. (2011), Circuit-level Soft-Error Mitigation, In: M. Nicolaidis (Ed), *Soft Errors in Modern Electronic Systems*, Springer, 2011.
- Yu, H., Nicolaidis, M., Anghel, L. & Zergainoh, N.E. (2011), Efficient Fault Detection Architecture Design of Latch-based Low Power DSP/MCU Processor, In *Proceedings, 16th IEEE European Test Symposium*, May 23-27, 2011, Trondheim, Norway.
- Zajac, P., Collet, J.H. & Napieralski, A. (2008), Self-Configuration and Reachability Metrics in Massively Defective Multiport Chips, in *Proc. 14th IEEE International On-Line Testing Symposium*, July 2008.

## STRUCTURAL CONSTRAINTS FOR THE CONSTRUCTION OF MULTI-STRUCTURED DOCUMENTS

PIERRE-ÉDOUARD PORTIER

*Université de Lyon, CNRS – INSA de Lyon – LIRIS UMR 5205*

*F-69621 France*

AND

Sylvie Calabretto

*Université de Lyon, CNRS – INSA de Lyon – LIRIS UMR 5205*

*F-69621 France*

**Abstract.** While are occurring the computer-mediated interactions for the weaving of relations between fragments of a documentary archive: structures appear, vocabularies emerge... Can programs be designed to help this effervescent creation not to diverge too quickly? One common solution is to rely on *a priori* well-defined and closed vocabularies (the so-called *ontologies*) from which the names being used to describe (annotate) and connect fragments are to be chosen. What can be done if such vocabularies aren't available? In other words: can a system be designed to allow the dynamic construction of vocabularies? We now propose a first version of such a system.

### 1. Introduction

We study the process of the construction of documents. We observe the emergence of documentary structures. This emergence relies on the creation of dimensions as sets of relations. We aim at providing computational mechanisms to assist the construction of dimensions. First of all, we introduce the notion of a non-trivial machine. By using a notion of computation seen as ordering, and by adopting a pragmatic point of view on the notion of meaning, we can redefine the objective as: programming mechanisms that could ease the circulation of information for the non-trivial machine.

### 2. Meaning and computation

J.V. Uexküll (1956), a father of ethology, developed a theory of meaning in order to explain in a unified way what he observed in many occasions on different kinds of animals: the same object placed in different environments can take a different meaning.

Thus he deduced that the qualities of an object are only perceptive attributes given by the subject with which they have a *connection*.

Furthermore, when G. Bateson (1972) wonders what it would mean for a computer to “think”, he comes to the conclusion that:

“What ‘thinks’ and engages in ‘trial and error’ is the man plus the computer plus the environment. And the lines between man, computer and environment are purely artificial, fictitious lines. They are lines across the pathways along which information or difference is transmitted.” p. 491.

Bateson tried to get rid of the subject/object dichotomy by considering systems described as networks of differences.

It links directly to a pragmatic view of meaning taken as an effect of the dynamic creation of relations. In (Saulnier and Longo, 2007), the idea of “conceptual frameworks” is introduced: meaning is to be found in the movements from one framework (or level of meaning) to another. Peirce’s concept of an *interpretant* is not far:

“A sign [...] creates in the mind of that person an equivalent sign, or perhaps a more developed sign. That sign which it creates I call the interpretant of the first sign.” (Peirce, 1897) (§228)

And the *meaning* would be this dynamic process of building an interpretant...

Finally, H. Von Foerster (2003) proposes a definition of computation as ordering. Ordering can be (i) a description of a given arrangement, or (ii) a re-arrangement of a (i). Moreover, he defines a non-trivial machine (Turing-like) as a machine for which the outputs depend on both the inputs and the state of the machine.

Thus, the frontiers of the considered non-trivial machine will include a computer and a user in an environment. This machine is in a dynamic state of producing orderings. “Meaning” is directly referring to this production. Indeed, the machine is powered by some desire (for example, the desire to explain a phenomenon) and the more the production of orderings fulfills the desire, the more meaningful the process is.

Our task is then to program some mechanisms that could ease the functioning of such a machine.

### 3. Construction of dimensions

#### 3.1. TREE CONSTRAINT

In the context of document engineering, what is commonly called “the problem of multi-structured documents” is the fact that elements of structures can be overlapping. Indeed, the most used formalisms for documents representation (first SGML, then XML) imply tree structures.

All of the models proposed to overcome this difficulty are centered on this tree/graph dichotomy. However, for each local event of two overlapping terms, those tend to belong to different dimensions or *levels of meaning*.

Thus, in the context of our multi-structured documents platform (Portier and Calabretto, 2010), each time an overlapping situation occurs with terms belonging to the same dimension, we offer the users the possibility to restructure the dimensions (see Figure 1).

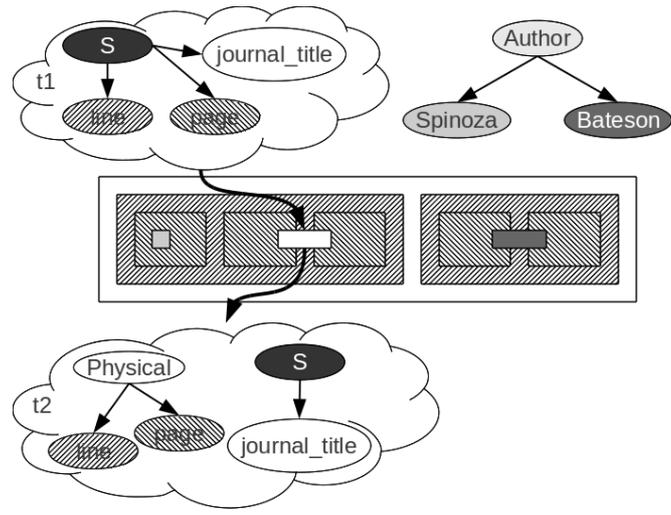


Figure 1. Formalization of user's knowledge when two terms of a same dimension overlap

STRAINT

Apart from the annotation of text intervals, relations are inter-weaved between heterogeneous fragments.

An essential part of the research on hyperstructures has created a notion of dimension. The zzstructure of T. Nelson (2004) for dimensional hypertexts is certainly one of the most relevant examples. The abstract function of a dimension is to group similar ways of weaving relations between fragments.

Indeed, a naïve graph-based representation doesn't offer appropriate synoptic views (see Figure 2). Thus, the dimensions provide clusters of relations that can compensate for this lack of synthesis by offering new kind of representations (see Figure 3).

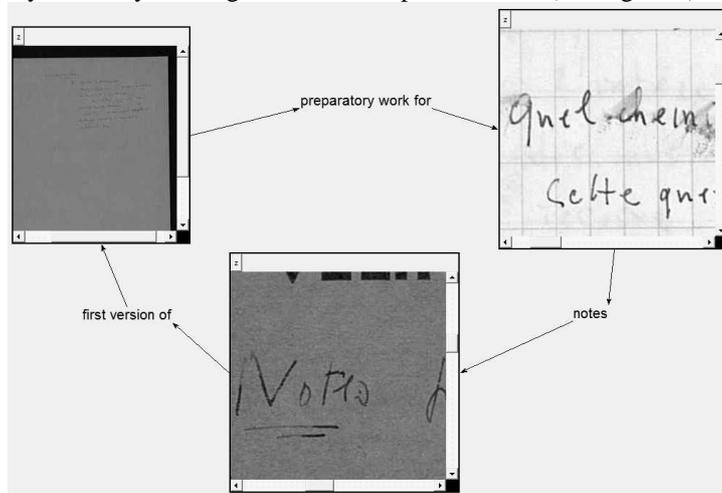


Figure 2. Illustration of a graph-oriented interface for the creation and the visualization of relations

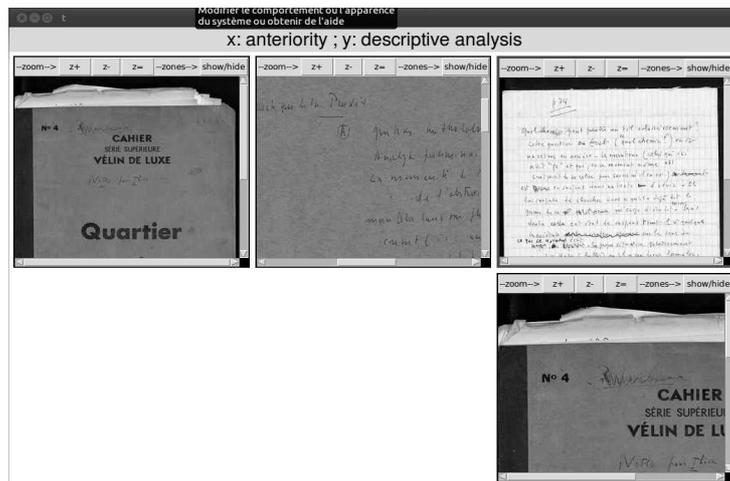
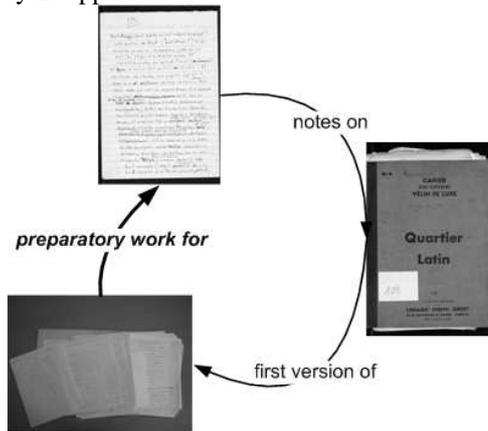


Figure 3. Illustration of a dimension

-based interface

In order to help the users in the process of creating dimensions, we are looking for a structural constraint whose violation is often meaningful and quite easy to dynamically detect.

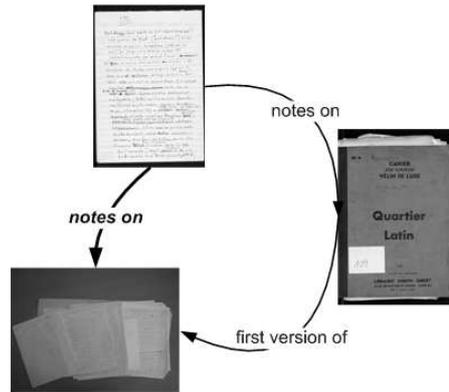
The acyclism constraint seems to be well adapted. Take for example the situation of Figure 4 where a user successively created two associations but when he adds a third relation, a cycle appears.



d
first version of
notes on
preparatory work for

Figure 4. After the free creation of some relations, a cycle appears within the “d” dimension

The user is advised to restructure the dimensions so as to remove the cycle (see for example Figure 5).



anteriority
first version of
preparatory work for

descriptive analysis
notes on

Figure 5. Formalization of structural users' knowledge after the automatic detection of a cycle within a dimension

4.

### Conclusion

This work is a first step towards a different point of view on computation seen as the construction of orderings by a non-trivial machine driven by a *desire* to explain some phenomenon. In such a configuration, new kinds of programs have to be developed in order to dynamically react to the user's actions by, for example, computing the appropriate times for helping the users to formalize their structural knowledge.

## Acknowledgements

We would like to thank the team of researchers from the Jean-Toussaint Desanti Institute for their collaboration during the development of this work.

## References

- Bateson, G. (1972). Steps to an ecology of mind. *The University of Chicago Press*.
- Nelson, T. H. (2004). A cosmology for a different computer universe: data model, mechanisms, virtual machine and visualization infrastructure. *Journal of Digital Information 5(1)*
- Peirce, C. S. (1897) Collected Papers of Charles Sanders Peirce 2. *Harvard University Press*, Cambridge
- Portier, P.-E., Calabretto, S. (2010). DINAH, a philological platform for the construction of multi-structured documents. In : *Proceedings of the 14th European conference on Research and advanced technology for digital libraries*, Glasgow, UK, p.364-375
- Saulnier, B., Longo, G. (2007). Le jeu du discret et du continu en modélisation : relativité dynamique des structures conceptuelles. In : *Intelligence de la complexité, épistémologie et pragmatique*. éditions de l'aube
- Von Foerster, H. (2003). Responsibilities of Competence. In Springer, ed., *Understanding understanding: essays on cybernetics and cognition*, p.191
- Von Uexküll, J. (1956). *Théorie de la signification*. Editions Denoël, Hambourg.

**(DIS-)TASTEFUL MACHINES?**

*Aesthetic Cognition and the Computational Turn in Aesthetics*

WILLIAM W. YORK

*Center for Research on Concepts and Cognition*

*Indiana University*

*512 North Fess Street*

*Bloomington, Indiana 47408-3822*

AND

HAMID R. EKBIA

*Center for Research on Mediated Interaction*

*Indiana University*

*1320 E. 10th Street*

*Bloomington, IN. 47405-3907*

**Abstract.** While aesthetics and cognition have traditionally been viewed as distinct from—even opposed to—one another, recent stirrings indicate the beginnings of an “aesthetic turn” regarding cognition. Does this, in turn, open up the possibility of a computational turn in the study of aesthetics? Can computational methods such as modeling and simulation be effectively brought to bear on something as mysterious and ineffable as aesthetic judgment? Or is “aesthetic cognition” a contradiction in terms? We explore these questions by focusing on the relationship between aesthetics and analogy-making, an area of cognition for which some research groundwork has already been laid. We will first offering some illustrative examples of this relationship, and then examine a group of computer models that have begun exploring mechanisms that may account for this relationship. Although rudimentary in their capabilities, these models point to a computational perspective for investigating not only the analogy–aesthetics relationship, but the processes underlying aesthetic cognition more generally.

**1. Introduction**

As Mark Johnson (2007) recently put it, “[A]esthetics is not just art theory, but rather should be regarded broadly as the study of how humans make and experience meaning” (p. 209). Aesthetic considerations factor into seemingly mundane everyday experience as well as in more exalted intellectual pursuits. Regarding the latter, Robert Root-Bernstein (2002) has used the term “aesthetic cognition” to refer to the “pre-logical, emotion-

laden, intuition-based feeling of understanding” (p. 62) that guides creative thought in science and mathematics.

In some quarters, the term “aesthetic cognition” might seem like a contradiction. There is a deeply rooted tendency to view the aesthetic and the cognitive as distinct from, if not opposed to, one another (Aiken, 1955). Yet recent stirrings from various quarters in cognitive science (e.g., Deacon, 2006; Norman, 2003) suggest that we are seeing the beginnings of an “aesthetic turn” in cognitive science.

## 2. A Computational Turn in Aesthetics?

Does this “aesthetic turn,” meanwhile, open up the possibility of a *computational turn* in aesthetics? Can the study of aesthetics be opened up to computational methods such as modeling and simulation? If so, how can they be effectively brought to bear on something as seemingly mysterious and ineffable as aesthetic sensibility? If not, what do we make of Root-Bernstein’s (2002) claim that “artificial intelligence will fail to provide insights into human thinking or model its capabilities until aesthetic cognition is itself understood sufficiently to be modeled and implemented by computers” (p. 75)?

Broadly speaking, there are two potential reactions to these questions. Optimistically, one might contend that fields such as cognitive science and artificial intelligence (AI) can—and, to some extent, already have—shed light on these questions, in part through the use of computer models, perhaps in combination with findings from neuroscience and experimental psychology. There is also the developing field of computational aesthetics (Hoenig, 2005). Despite its somewhat different emphases—which range from image-processing techniques to computer-generated art to formal analysis of artworks—the growth of this new field offers further evidence of the potential relevance of computation to aesthetics (and vice versa).

In turn, skeptics might reply that longstanding problems in aesthetics have remained unsettled for a reason: There may simply be limits to what we can understand when it comes to matters of judgment, sensibility, and taste (Weizenbaum, 1976). To explain aesthetic sensibility would seem to involve specifying, formalizing, or mechanizing those same intuitive processes that have been *defined* as unspecifiable, unformalizable, or non-mechanizable (e.g., Polanyi, 1981; Dreyfus, 1992). This debate between optimists and skeptics is ongoing, encompassing other areas of human cognition and behavior; in particular, it has been framed around various theories and models in artificial intelligence (Eklbia, 2008). Is there a meaningful way to resolve, or at least advance, this debate?

## 3. Analogy-Making as Aesthetic Cognition

The perceptual and (especially) the aesthetic dimensions of analogy-making have been downplayed in much research on analogy within cognitive science and AI, where the focus has instead been on “analogical reasoning” (e.g., Winston, 1980). Yet analogy is not coextensive with reasoning, and the idea that analogy-making involves an aesthetic component does have some precedence. For example, in the program Copycat—a model of analogy-making in the microdomain of letter strings (e.g., “If **abc** is changed to **abd**,

then how should **kkjji** be changed?”)—the “computational temperature” at the end of a run can be construed as a sort of aesthetic evaluation of the program’s answer (Mitchell 1993). Copycat’s successor, Metacat, is able to compare different answers to a given analogy problem—say, **kkjjhh** and **kkjjij** in response to the example given above—on the basis of three largely aesthetic dimensions: *uniformity*, *abstractness*, and *succinctness* (Marshall 1999).

Likewise, the idea that aesthetic sensibility involves an ability to perceive and appreciate analogies has also been noted before. For example, Koestler (1964) refers to the “hidden analogies” that inform the creative process in science, art, and humor. Arnheim (1969) discusses the role of analogy in the perception and grouping of visual forms, including what might be called “visual rhymes.” Similar types of analogical mappings can be identified in the plot structures of films, novels, and other narrative forms. Meanwhile, the role of aesthetic factors in science and mathematics has also been explored (e.g., Papert, 1988; Sinclair, 2004), further highlighting the connection between aesthetic sensibility, insight, perception, and analogy. Finally, computer models such as Letter Spirit (Rehling, 2001) have explored the role analogy in the more traditionally aesthetic realm of alphabetic font (or grid font) design.

#### 4. Open Questions

Models such as Copycat and Letter Spirit suggest a potentially rewarding perspective for investigating not only the analogy–aesthetics relationship, but the processes underlying aesthetic cognition more generally. But to what extent can such computational approaches ultimately contribute to this joint understanding? What are the strengths (and limits) of computer models that aim to simulate the processes of analogy-making and aesthetic judgment in human beings? Finally, is there potential for common ground between cognitive science/AI and the growing field of computational aesthetics?

#### Acknowledgements

Thank you to Helga Keller (R.I.P.) for her tireless support over the years.

#### References

- Aiken, H. D. (1955). Some notes concerning the cognitive and the aesthetic. *The Journal of Aesthetics and Art Criticism*, 13(3), 378–394.
- Arnheim, R. (1969). *Visual Thinking*. Berkeley: Univ. of California Press.
- Deacon, T. (2006). The aesthetic faculty. In M. Turner (Ed.), *The Artful Mind: Cognitive Science and the Riddle of Human Creativity* (pp. 3–20). Oxford: Oxford Univ. Press.
- Dreyfus, H. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, Mass.: MIT Press.
- Ekbia, H. R. (2008). *Artificial Dreams: The Quest for Non-Biological Intelligence*. Cambridge, U.K.: Cambridge Univ. Press.

- Hoenig, F. (2005). Defining computational aesthetics. In L. Neumann, M. Sbert, B. Gooch, and W. Purgathofer (Eds), *Computational Aesthetics 2005: Eurographics Workshop on Computational Aesthetics in Graphics, Visualization, and Imaging* (pp.13–18).
- Johnson, M. (2007). *The Meaning of the Body: Aesthetics of Human Understanding*. Chicago: Univ. of Chicago Press.
- Koestler, A. (1964). *The Act of Creation*. New York: MacMillan.
- Marshall, J. (1999). *Metacat: A Self-Watching Cognitive Architecture for Analogy-Making and High-Level Perception*. Doctoral dissertation, Indiana Univ., Bloomington.
- Mitchell, M. (1993). *Analogy-Making as Perception: A Computer Model*. Cambridge, Mass.: MIT Press.
- Norman, D. (2003). *Emotional Design: Why We Love (or Hate) Everyday Things*. New York: Basic Books.
- Papert, S. (1988). The mathematical unconscious. In J. Wechsler (Ed.), (1988), *On Aesthetics in Science* (pp. 105–120).
- Polanyi, M. (1981). The creative imagination. In D. Dutton & M. Krausz (Eds.), *The Concept of Creativity in Science and Art* (pp. 91–108). The Hague, Netherlands: Nijhoff.
- Rehling, J. A. (2001). *Letter Spirit (Part Two): Modeling Creativity in a Visual Domain*. Doctoral dissertation, Indiana Univ., Bloomington.
- Root-Bernstein, R. S. (2002). Aesthetic cognition. *International Studies in the Philosophy of Science*, 16(1), 61–77.
- Sinclair, N. (2004). The roles of the aesthetic in mathematical inquiry. *Mathematical Thinking and Learning*, 6(3), 261–284.
- Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco: W. H. Freeman and Co.
- Winston, P. H. (1980). Learning and reasoning by analogy. *Communications of the ACM*, 23(12), 689–703.

# **Track VII: Social Computing**

## The social and its political dimension in software design

### *A Socio-Political Approach*

DORIS ALLHUTTER

*Austrian Academy of Sciences, Institute of Technology Assessment  
Strohgasse 45, 1030 Vienna*

**Abstract.** Recent debates in philosophy and computing and science and technology studies address the prolongation of the social in technical design and development and thus the question of discursive performativity. Applying a wider conception of *the social* than usually referred to in design research, I present an initial elaboration of a socio-political approach to software design. This approach is based in discourse theory, deconstructivism and ‘new materialism’ and focuses on the reproduction of power by tracing the performativity of hegemonic societal discourses and their co-materialization with (normative) technological phenomena. Making use of Karen Barad’s material-discursive account of performativity, I argue that a socio-political approach to software design needs to take into account the ‘intra-action’ of material phenomena with reconfigurings of power relations in intertwined epistemic and everyday work practices. The objectives of this endeavour are, *first*, to ask and make negotiable who (in/formal hierarchies) and what (discursive hegemonies) is given normative power in design processes on the basis of which social and technological imaginaries; *second*, to investigate and, to some extent, try to make tangible how these—mostly unconscious—normative enactments co-materialize with material phenomena or relations; and *eventually*, to elaborate on how to widen human agency by opening spaces for maneuver or trading zones when taking account of the agency of human/non-human assemblages or material-discursive re-configurations of the world.

Recent debates in philosophy and computing and science and technology studies have expanded the question of the prolongation of the social in technical design and development by taking into account the concept of discursive performativity. Inspired by this discussion and applying a wider conception of *the social* than usually referred to in research on the development of computational artifacts, I present an initial elaboration of a socio-political approach to software design. This socio-political approach connects to the notion of ontological politics (see Mol, 1999) and is based in discourse theory, deconstructivism and ‘new materialism’. It focuses on the reproduction of power by tracing the performativity of hegemonic societal discourses and their co-materialization with (normative) technological phenomena.

Karen Barad’s (2007) materialistic elaboration of the concept of performativity shifts the focus from a linguistic and discursive account of performativity, which is linked to the paradigm of the co-construction of society and technology, to the notion of co-materialization. She criticizes earlier approaches to processes of materialization (as for example introduced by Butler and Foucault) that centre on the question of ‘how discourse comes to matter’. Barad suggests that their focus on the social constructedness

of bodies/materiality in fact neglects the question of ‘how matter comes to matter’ and puts an equal focus on the material dimensions of agency.

In my previous work, Donna Haraway’s account of ‘embodied, situated practices’ and Judith Butler’s concept of discursive performativity have inspired me to investigate software design processes as entangled practices informed by technological concepts and hegemonic societal discourses as much as by professional self-conceptions of developers and related workplace politics (see Allhutter 2011). Barad’s materialistic move that resulted in her elaboration of ‘agential realism’ can add to such a perspective on software design in that it conceptually takes into account the agency of materiality or material phenomena (see also Velden and Mörtberg, 2011). Still open remains the question of how to make use of a material-discursive account of performativity in applied design research.

In this respect, I suggest that it makes sense to reconstruct the journey of two crucial concepts—‘agency’ and ‘materialism’—that have been travelling between disciplines and research fields: While questions of the agency of artifacts and human/non-human (re-)configurations have intensively been discussed in studies of science and technology since the early 1980ies (Callon, Latour, Law, Haraway), only recently political science scholars such as Jane Bennet (2010), Diane Coole and Samantha Frost (2010) have begun to integrate this strand of theory to rethink concepts of political agency and to rework the notion of materialism, now discussed as ‘new materialisms’.

On this background, I argue that a socio-political approach to software design practice and theory needs to take into account the ‘intra-action’ of material phenomena with reconfigurings of power relations (normativity and societal hegemonies) in intertwined epistemic and everyday work practices. My objective of elaborating such a socio-political approach based on a material-discursive account of performativity is threefold:

First, the aim is to ask and make negotiable who (in/formal hierarchies) and what (discursive hegemonies) is given normative power in design processes on the basis of which social and technological imaginaries (e.g. re-enactments of societal differences and epistemic dichotomies); second, to investigate and, to some extent, try to make tangible how these—mostly unconscious—normative enactments co-materialize with material phenomena or relations (that are e.g. development methods, processes, artifacts); and eventually, to elaborate on how to widen human agency by opening spaces for maneuver or trading zones (Allhutter and Hofmann, 2010) when taking account of the agency of human/non-human assemblages or material-discursive re-configurations of the world.

## References

- Allhutter, D. (2011). Mind Scripting: A Method for Deconstructive Design. *Science, Technology & Human Values*, OnlineFirst March 13, 2011.
- Allhutter, D. & Hofmann, R. (2010). Deconstructive Design as an Approach to opening Trading Zones. In: J. Vallverdú (ED), *Thinking Machines and the Philosophy of Computer Science: Concepts and Principles* (pp. 175–192). Hershey: IGI Global.
- Barad, K. (2007). *Meeting the Universe Halfway: Quantum physics and the entanglement of matter and meaning*. Durham and London: Duke University Press.

- Bennet, J. (2010). *Vibrant Matter: A political ecology of things*. Durham and London: Duke University Press.
- Coole, D. & Frost, S. (2010). *New Materialisms: Ontology, Agency, and Politics*. Durham and London: Duke University Press.
- Mol, A. (1999). Ontological Politics: a Word and Some Questions. In: J. Law and J. Hassard (EDS), *Actor Network and After*. (pp. 74–89). Oxford and Keele: Blackwell and the Sociological Review.
- Velden, M. van der & Mörtberg, C. (2011). Between Need and Desire: Exploring Strategies for Gendering Design Science, *Technology & Human Values*, OnlineFirst March 13, 2011.

## A SOCIAL EPISTEMOLOGICAL APPROACH FOR DISTRIBUTED COMPUTER SECURITY

Steve Barker  
*Department of Informatics*  
*King's College London*

**Abstract.** We present a social epistemological approach, for treating an aspect of computer security, which allows for multiple testifiers to contribute propositional attitude reports to a community repository of testimonial knowledge and for users to adopt a range of epistemic positions for deciding what constitutes justified belief in different contexts.

### 1. Introduction

We discuss a key epistemological aspect of the distributed access control (DAC) problem: in large, distributed computer systems, like the Internet, how can a decision be rendered on whether a requester of access to a resource is authorised to perform an action on the resource if what is *known* by the decision-maker about a requester is “incomplete”? (And it is computationally too expensive for the decision-maker to exhaustively search for all of the knowledge it (ideally) requires on the requester.)

Rather than simply rejecting the access request on the basis of the incompleteness of its knowledge, the putative solution to the DAC problem is for the decision-maker to accept the assertions of some individual, ultimately *trusted* testifier who “speaks for” the requester and in so doing enables the decision-maker to determine whether the requester is authorised to perform a requested action on a resource. The notion of an ultimately trustworthy source of epistemic warrant assumes that a *foundationalist* (Bonjour 1985) position on knowledge/justification applies in the DAC case; there is no infinite justificational regress because what the trusted source asserts is so.

In Section 2 of this abstract, we suggest an alternative, social epistemological approach to the DAC problem. In Section 3, we draw conclusions.

### 2. An Alternative Approach to the DAC Problem

We argue for a community-based approach to testimonial warrant and for testifiers making assertions of their *propositional attitudes* (Russell 1905) via a community-based repository, which is a store of triples  $(s, \alpha, p)$  such that  $s$  is a source of assertions in a community of sources  $\Sigma = \{s, s_1, \dots, s_n\}$  of testimonial warrant,  $p$  is a proposition, and  $\alpha$  is a propositional attitude that a source in  $\Sigma$  has in relation to  $p$ .

We note that  $p$  may be an atomic proposition or an arbitrary logical formula, we restrict attention to the doxastic attitudes “believes” and “disbelieves”, and we interpret a source as suspending belief on  $p$  if it makes no assertion of  $p$  to the community repository. The triples  $(s_i, \alpha, p)$  represent *that*-clauses, e.g.,  $s_i$  believes that  $s_j$  is “bad debtor”. Typically, in the DAC scenario, the assertions are on a requester’s reputation, e.g., for being a “bad debtor”; the categories of requesters to be used are community determined. In the context we assume, authorisation depends on the assignment of a requester to a category, e.g.,  $S$  is authorised to perform some action on a resource iff  $S$  is categorised as a “good trader” (say). We suggest that what we propose is appropriate for addressing the DAC problem in that it recognises the need for knowledge construction by a division of epistemic labour, it allows for justified belief to be community constructed (which we hold to be more reliable than exclusively using individual, foundational sources of testimonial knowledge) and it recognises that, in the context of interest, “truth” is appropriately held to be relative to a community.

It is open to decision-makers to decide what methods of computation to use, with the community repository, for them to have justified beliefs for deciding on authorisation requests. A decision-maker may simply accept that the propositional attitude  $\alpha$  holds in relation to  $p$  if some specific source  $s \in \Sigma$  expresses that directly. However, this is far from being the only option. A decision-maker may, for example, accept that  $\alpha$  holds in relation to  $p$  because some, non-specific member of  $\Sigma$  asserts that or all members of  $\Sigma$  assert that or it is the “majority view” (variously interpreted) of members of  $\Sigma$  that  $\alpha$  holds in relation to  $p$ . Moreover, more complex requirements may be expressed in more expressive logic languages, e.g., an acceptor may accept that  $\alpha$  applies in relation to  $p$  if some  $s_i \in \Sigma$  asserts that and no source in  $\Sigma$  disbelieves  $p$ . It is important to note that we allow individual decision-makers to decide on what constitutes evidence for them “knowing” that an authorisation holds, that the knowledge for this is socially constructed, and that different forms of inferential knowledge will be applicable for decision-making in different contexts (cf. DeRose 1992).

In the evidentialist framework that we adopt (Feldman and Conee 1985), we say that: *a decision-maker  $\gamma$  is justified in adopting the assertion by  $s \in \Sigma$  that the propositional attitude  $\Sigma$  holds in relation to the proposition  $p$  at the time  $t$  iff the attitude  $a$  on  $p$  is entailed by some computational method that  $\gamma$  justifiably holds to be reliable for this entailment at the time  $t$  from the evidential sources that  $\gamma$  justifiably holds to be sufficiently authoritative for the purpose of making the inference that  $a$  holds on  $p$  according to  $s$  at  $t$ .*

Evidentialist-based interpretations of a variety of epistemic positions will be adopted in practice. It follows that we do not argue that foundationalism is not a meaningful epistemic position to adopt in the DAC context. Rather, we suggest that different epistemic positions (e.g., foundationalist, Haackean foundherentist, etc.) will apply in different contexts. It is the emphasis on a plurality of epistemic positions that is distinctive about our approach.

### 3. Conclusions

We critically assessed the foundationalist epistemic position that has hitherto been assumed in treating the DAC problem. We then argued for a social epistemological alternative, which accommodates propositional attitude reports, community-based testimonial assertions and the flexible use of a range of methods for producing inferential knowledge.

In future work we intend to consider repositories that maintain a history of propositional attitudes and the epistemic issues that arise.

### **References**

- Bonjour L. (1985). *The Structure of Empirical Knowledge*. Harvard University Press.
- DeRose, K. (1992). Contextualism and Knowledge Attributions, *Philosophy and Phenomenological Research*, 52, pp. 913-929.
- Feldman R. and Conee E. (1985). Evidentialism, *Philosophical Studies*, 48, pp. 15-34.
- Russell, B. (1905). On Denoting, *Mind*, 14, pp. 479-93.

## TRUST, POWER, AND INFORMATION TECHNOLOGY

MARK COECKELBERGH

*University of Twente*

*Department of Philosophy, P.O. Box 217, 7500 AE Enschede, The Netherlands, E-mail m.coeckelbergh@utwente.nl*

**Abstract.** This paper offers a preliminary discussion of the relation between trust, power, and information technology. It also explores some implications for ethics and politics of information technology.

### 1. Introduction

In recent years the issue of trust has received much attention in ethics and philosophy of information technology. For instance, there is work on e-trust and on-line trust: some argue against e-trust (for example Nissenbaum 2001), while others are more optimistic about trust in digital contexts (Taddeo 2009, 2010a, 2010c, Turilli et al 2010). Furthermore, in the field of social epistemology there is work on trust and knowledge (Simon 2009, Taddeo 2010b), and people working in the virtue ethics and phenomenological tradition have developed a notion of 'implicit' trust (Ess 2010, Carusi 2009).

While this attention to trust has produced insightful work relevant to both philosophers and computer scientists who try to model trust, there is little or no attention to relations between trust, power, and information technology. This paper is a preliminary attempt to explore this relation. First I will clear the ground by making a claim regarding the epistemology of trust (I will need this later), then I will make two claims about the relation between trust and power: (1) trust presupposes power relations and (2) trust creates power relations.

This analysis will allow me to make some suggestions about the implications for ethics and politics of information technology.

### 2. Trust, Knowledge and Transparency

Although it is true that trust can emerge in uncertain and risky on-line environments and that in one sense trust promotes transparency, as Turilli and others have argued (Turilli et al 2010), there is also a sense in which (a) trust can only exist under conditions of uncertainty and (b) transparency destroys trust.

In order to develop these claims, we must challenge the rationalist-contractarian assumption entertained in Taddeo's work, that e-trust cannot appear a priori, but depends on the assessment of trustworthiness by a rational (artificial) agent (Taddeo 2010c). A phenomenological notion of trust, by contrast, involves a sort of a priori, implicit form of trust. This form of trust flourishes only in environments characterized by incomplete certainty, knowledge and transparency. If there was complete uncertainty, complete lack of knowledge, and no transparency at all, we would have no basis for trust. On this point rationalist-contractarian models are right. If, however, if there was complete knowledge, complete certainty, and full transparency, there would be no need for trust; the problem would not arise in the first place.

This suggests that if political movements aim for total, absolute transparency (e.g. Wikileaks), they risk to destroy trust, which must be situated 'in between' the epistemic absolutes identified.

However, this is a claim about knowledge; what about trust with regard to action?

### **3. Trust and Power (1)**

If trust is not entirely freely decided by rational agents, but presupposed in social relations, then we need to discuss how prior social relations, understood as power relations, shape trust. There are a priori dependencies that enable but also constrain agency with regard to trust. In a particular social network, I 'have' to trust some others and indeed some technologies (e.g. software) since, and to the extent that, I am dependent on them for the very practice I am engaged in. In any social network, I am dependent on some key, powerful actors and technologies which I 'have' to trust *because* they are powerful. This means limits my agency with regard to trust. Power relations – relations with others and with technologies – already shape trust 'before' any decision or deliberation about trust is made.

If this is true, it does not only set limits to efforts to model and implement trust in artificial networks, it is also relevant for ethical-philosophical analysis of trust in digital environments 'inhabited' or 'crawled' by both humans and artificial agents. In the digital age, trust crucially depends on power exercised by the 'architects', 'providers' and 'webmasters' of the social-technological networks that form and transform our interactions and practices (including academic practice).

But how did these social actors become powerful in the first place? Does this analysis preclude agency altogether?

### **4. Trust and Power (2)**

Even a strictly rationalist-contractarian approach to trust must acknowledge that trust, 'decided' upon by rational agents, creates power relations and generates its own normativity with regard to humans and their artificial cooperants.

If an agent A says 'I trust you' to an agent B, this does not only create expectations A has about B's future actions, but also involves a delegation of (discretionary) power from A to B. In addition, and this is the normative aspect, A makes B responsible. If A trusts B to do something, then A holds B responsible for doing that. In particular, if B

decides to do otherwise (trust presupposes that B has this space of freedom), then B has to provide reasons to A, explain why (s)he did not do what A expected him or her to do. Trust is violated if no good reasons are given by B.

This analysis of relations between trust, power, and normativity is relevant for 'horizontal' social relations, but also for the 'vertical' relation between individuals and the state. This works both ways:

(1) an individual A may trust state B, which implies that A delegates power to B to do something and that B becomes responsible. A's trust can then be violated by B if B fails to do this and if fails to give good reasons for not doing it.

(2) state A can trust its citizens B (not) to do something, that is, hold B responsible, and B can violate this trust.

## 5. Conclusion

I conclude that this framework, which tolerates and employs both rationalist-contractarian and phenomenological approaches, reveals a lacuna in the present literature and allows us to analyze and discuss the power dimension of issues in social epistemology, information ethics and philosophy of information.

For example, in the Wikileaks case, there seems to be a clash between on the one hand a vertical 'delegation' model, which creates the possibility of trust under conditions of uncertainty, and on the other hand a model that aims at transparency, attempts to provide complete knowledge, and seeks to abolish the vertical delegation relation – and thereby abolishes trust in the sense discussed above.

Of course this analysis does not exhaust the many interpretations of the word 'trust' used in the literature. And perhaps a tension remains between rationalist- contractarian and phenomenological approaches. Furthermore, neither power nor trust should be our only concern in ethics and politics of information technologies. However, I hope this exploration of the relation between trust, power, and information technologies can contribute to the expanding research on trust and information technology.

## References

- Carusi, A. (2009). Implicit Trust in the Space of Reasons: A Response to Justine Pila. *Journal of Social Epistemology* 23(1), 25-43.
- Ess, C. 2010. Trust and New Communication Technologies. *Knowledge, Technology, & Policy* 23(3-4), 287-305.
- Nissenbaum, H. (2001). Securing Trust Online: Wisdom or Oxymoron. *Boston University Law Review* 81(3), 635-664.
- Simon, J. (2009). Webs of Trust and Knowledge: Knowing and Trusting in the World Wide Web. In: *Proceedings of the WebSci'09: Society On-Line*, 18-20 March 2009, Athens, Greece.
- Taddeo, M. (2010a). Trust in Technology: a Distinctive and a Problematic Relation. *Knowledge, Technology and Policy* 23 (3-4), 283-286.
- Taddeo, M. (2010b). An Information-Based Solution for the Puzzle of Testimony and Trust. *Social Epistemology* 24(4), 285-299.
- Taddeo, M. (2010c). Modelling Trust in Artificial Agents: A First Step toward the Analysis of e-Trust. *Minds and Machines* 20(2), 243-257.

- Taddeo, M. (2009). Defining Trust and E-trust: Old Theories and New Problems. *International Journal of Technology and Human Interaction* 5(2), 23-35.
- Turilli, M, Vaccaro, A., & Taddeo, M. (2010). The case of on-line trust. *Knowledge Technology and Policy* 23(3-4), 333-345.

## **THE BENEFITS OF SOCIAL THEORY FOR MODELLING STABLE ENVIRONMENTS OF SYSTEMIC TRUST WITHIN MULTI AGENT SYSTEMS**

DIEGO COMPAGNA

*University of Duisburg-Essen, Institute of Sociology*

*Lotharstr. 65 (LE 643), 47057 Duisburg*

### **1. Modelling Stable Environments of Systemic Trust within Multi Agent Systems**

Trust is often discussed on the micro-level of individuals or discrete entities; instead I would like to stress the benefits of systemic trust that could be seen as a form of mediated trust between entities. Based on the proposition of the 'Homeostatic Feedback Loop' by Anthony Giddens a stable social environment can be modeled for Multi Agent Systems (MAS). The goal of this Model is on the one hand trust is build as a non-intended effect on the systemic level from which on the other hand all participating entities take benefit: The outcome is an auto-sustaining framework; or a homeostatical systemic state. In this model trust emerges as the result of non intended effects of distinct actions between different Agents that could be described as a functional cooperation.

The specific characteristic of the Casual Feedback Loop - the core proposition within the notion of a duality of structure (Giddens 1984) - could be very useful for a MAS architecture that enfolds a stable environment (Compagna 2009). The main assumption behind the concept of the duality of structure is that actions and the framework of these actions are organized recursively, or in terms of the social system theory in the modus of an autopoietic sustainment (Giddens 1991). Within such an environment of mutual but non-intended functionality the value of trust become an emergent value or a non-intended outcome. Based on an early Paper of Castelfranchi/Conte (1992) different kinds of cooperation could be described: Non-Intended, Intentional, Out-Designed and Functional. Functional Cooperation is described as the best way to establish a fruitful and stable cooperation between agents. This type of cooperation could be related and captured as well as further conceptualized very well with the Theory of Structuration.

The model I would like to present - by combining the above mentioned propositions - consists in the mutual goal for the involved agents of an action-framework that is functional for them although this is not directly intended by their intentionally motivated actions. Although this model claims to explain and accomplish a stable framework for MAS it could be transferred to a Human-Agent set-ting in which by non-intended effects a stable interaction framework emerges that provides a favorable context for mutual system trust.

## References

- Castelfranchi, Cristiano & Conte, Rosaria (1992). Emergent functionality among intelligent systems. Cooperation within and without minds. In: *AI & Society* 6 (1), S. 78-87.
- Compagna, Diego (2009). *Sozionik und Sozialtheorie. Zum Beitrag soziologischer Theorien für die Entwicklung von Multiagentensystemen.* (1. Aufl.) Saarbrücken: VDM Verlag.
- Giddens, Anthony (1984). *The constitution of society. Outline of the theory of structuration.* (1. Aufl.) Cambridge: Polity Pr. [u.a.].
- Giddens, Anthony (1991). *Structuration theory. Past, present and future.* In: Bryant, Christopher G.A. / Jary, David (Hg.): *Giddens' theory of structuration. A critical appreciation.* (1. Aufl.) London [u.a.]: Routledge. (S. 201-221)

## COMPUTER NETWORKS AND THE PHILOSOPHY OF MIND

### *A Social Mind – Networked Computer Analogy*

ISTVAN DANKA

*Department of Philosophy, University of Leeds  
Leeds, LS2 9JT, United Kingdom*

In the last few decades, computer analogies of the mind have dominated several central fields of the philosophy of mind. The leading versions of the 'mind – computer' analogy are based on the Interface Model of the Mind (with Putnam's phrase), claiming that the mind of an individual is analogous to a computer with an interface connection to its environment. As opposed to this, I shall develop a Network Model of the Mind, based on an analogy between the socially extended mind and a computer network, according to which social relations and semantic content of the WWW are analogously structured. In accordance with Clark and Chalmers' extended mind hypothesis, I shall argue that there are active constituent parts of mental processes that are located externally to the mind of an individual, just as there are semantic contents external to individual computers.

A network model of the mind is the opposite of the interface model in the following sense. The interface model rests on the (Cartesian-inspired) assumption that there is a surface on which the mind interacts with its environment. For a social *externalist* the mind is extended over the limits of the body and hence no "surface" of the individual can be drawn. For a *social externalist*, mental processes are more plausibly understood as social activities among interlinked individuals. In either case, it makes no sense alluding to any interface. For a network model, what is essential in the structure of mental contents is not separation but connection. Hence, it explains the mental in terms of connections among mental contents in the minds of different individuals.

At least two significant versions of the 'social mind – networked computer' analogy can be developed. On the one hand, one can argue for an analogy between socially embedded individual minds and networked computers. In this case, the connections have to be understood as physical connections among computers (i.e., the internet) on the one hand, and socially connected individual humans (social networks) on the other. The second version is philosophically more interesting though. Namely, an analogy can be drawn between semantic content on the net (WWW) on one hand, and mental content structured socially on the other hand. This analogy demonstrates that mental contents cannot be individually located in our heads since, analogically, semantically significant units of the content are not necessarily contained by the server but they are often spread over multiple machines (e.g. cookies).

Regarding the connections among mental contents, I shall distinguish three structurally different models of the individual mind in terms of the relations among mental contents. First, centralised (Cartesian/Kantian) views argue that there is a centre

of mental content (the soul, the mind, the Self, etc.), to which all mental contents are (directly or indirectly) connected. Second, non-centralised (behaviourist/physicalist) views claim that no centre of mental contents is provided; the best model for the relations among mental contents is a random graph. Third, de-centralised models (e.g. Quine) claim that there is a difference between central and peripheral mental contents; though no clear distinction can be made between the contingent and the necessary, a gradual account of more and less central contents can be provided.

In parallel, there are three main models of the social relations among mental contents. Those who accept centralised models of the individual mind will most probably follow a multi-centred view of the social, claiming that mental contents constitute many centres of individual minds connected to each other randomly. (A logically possible alternative to this would be arguing that there is a centre of the social as well, but no serious attempt has been made in order to support such a view.) Holders of non-centralised models of the individual can apply their random graph set to the social, claiming an equal distribution of socially explained connections among mental contents. Finally, defenders of the de-centralised view claim that there are socially more and less central contents and even if there is no single centre of the social, several hubs can be identified.

Analysing different approaches to how semantic content on the internet is organised, I shall develop a topology of networked-based relations among mental contents and argue for a de-centralised network model of the social mind, based mostly on an analogy with A-L. Barabási's research on the topology of the internet. While doing so, I shall allude to (1) the unequal distribution of links on the internet (the "rich get richer" phenomenon), (2) the impossibility of complex networks' being centralised ("the winner does not take all"), and some differences between inbound and outbound links regarding the semantic significance of web pages. Based on these, I shall argue for a de-centralised network model for the social mind, following an analogy between the structure of the content on the WWW and a graph theoretically equivalent model of the mind to Quine's gradual approach between the central and the peripheral. However, there is a slight modification in my own version. From the network analogy it follows that the building of knowledge is not hierarchical, though it is also not an evenly distributed random model of connections among items. However, the least connected items are not connected to gradually more connected items while reaching highly connected items. On the contrary: they are mostly directly connected to "central" hubs. Therefore, a spatial metaphor of 'central vs. peripheral' is misleading.

All the same, it can also be argued that even though the (physical) structure of the internet and the (semantic) structure of the WWW are analogous (and hence are the structure of mental contents and that of social relations), the connection between the two is contingent. Since from the analogy it follows that a multi-centred view of the social mind is incompatible with the actual structure of the semantic on the web, on the supposition of the analogy, no item of mental contents can be located in individuals. Hence, no interface can be identified. If so, the 'social mind – networked computer' analogy may serve as a useful weapon of social externalists.

## AGENT BASED MODELING WITH APPLICATIONS TO SOCIAL COMPUTING

Gordana Dodig Crnkovic  
*School of Innovation, Design and Engineerin,*  
*Mälardalen University, Sweden*  
*gordana.dodig-crnkovic@mdh.se*

### 1. Extended Abstract

Even though computers were invented primarily to automatize calculations, already Licklider and Taylor (1968) emphasized the importance of the computer as a communication device, with consequent shared knowledge and community-building.

There are two different approaches to social computing, (Wang et al. 2007), one with the strong emphasis on *technological, computing side* and the other centered on *human, social aspect*. Present analysis will be focusing the first kind of social computing, a computational approach to modeling of social interactions, including the development of their supporting information and communications technologies. The main tools are *simulation techniques* used in order to facilitate the study of society and to support decision-making policies, helping to analyze how changing policies affect social, political, and cultural behavior (Epstein, 2007).

Social computing is radically changing the character of human relationships worldwide (Riedl, 2011). Instead of maximum 150 connections prior to ICT (Dunbar, 1998) present social computing easily leads to networks of several hundred of contacts. It remains to understand what type of society will emerge from such massive “long-range” distributed interactions instead of traditional fewer and deeper short-range ones.

As in the process information overload on individuals is steadily increasing, social computing technologies are moving beyond social *information processing* toward *social intelligence*, (Zhang et al. 2011) (Lim et al. 2008) (Wang et al. 2007), which brings an additional level of complexity.

Social computing with the focus on *social* is a phenomenon which enables *extended social cognition*, while the social computing with the focus on *computing* is about computational modeling and *new paradigm of computing*. I will focus on the agent-based social simulation (ABSS) as a generative computational approach to social simulation defined by the interactions of autonomous agents whose actions determine the evolution of the system, as applied in artificial life, artificial societies, computational sociology, dynamic network analysis, models of markets, swarming (including swarm robotics) (Antonelli and Ferraris 2011), (Chai et al., 2010). As Gilbert (2005) rightly points out, novelty of agent based models (ABMs) “offer the possibility of creating ‘artificial’ societies in which individuals and collective actors such as organizations could be

directly represented and the effect of their interactions observed. This provided for the first time the possibility of using experimental methods with social phenomena, or at least with their computer representations; *of directly studying the emergence of social institutions from individual interaction.*" ABMs are very useful computational instruments but they should not be taken as "reality" even though simulations with their realistic graphical representations suggest their being "real". Process of modeling and simulation is complex and many simplifications and assumptions must be made which always must be justified for each application. (Gilbert and Troitzsch 2005) Grimm and Railsback 2005) (Axelrod 1997)

ABMs in general are used to model complex, dynamical adaptive systems (Breiger et al. 2003). The interesting aspect in ABMs is the micro-macro link (agent-society). Multi-Agent Systems (MAS) models may be used for any number (in general heterogeneous) entities spatially separated by the environment which can be modeled explicitly. Interactions are in general asynchronous which adds to the realism of simulation. (Miller and Page 2007) (Schuler 1994)

Social computing represents a new computing paradigm which is one sort of the natural computing, often inspired by biological systems such as e.g. swarm intelligence, evolutionary computation or artificial immune systems. In my analysis I will present different paradigms of computation including social computing and modeling of cognitive agents in the info-computational framework (Dodig-Crnkovic 2011) (Dodig-Crnkovic and Müller 2009).

## References

- Antonelli C. and Ferraris G. (2011) "Innovation as an Emerging System Property: An Agent Based Simulation Model", *Journal of Artificial Societies and Social Simulation JASSS* 14 (2) 1, <http://jasss.soc.surrey.ac.uk/14/2/1.html>
- Axelrod, R. (1997). *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton: Princeton University Press.
- Breiger R., Carley K. and Pattison P. (2003) *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, Nat'l Academies Press.
- Chai S-K, Salerno J. and Mabry P. L. (eds.) (2010) "Advances in Social Computing: Third International Conference on Social Computing, Behavioral Modeling, and Prediction", SBP 2010, Bethesda, MD, USA Springer-Verlag: Berlin.
- Dodig-Crnkovic G. (2011) "Significance of Models of Computation from Turing Model to Natural Computation." *Minds and Machines*, DOI 10.1007/s11023-011-9235-1. Special issue on Philosophy of Computer Science; R. Turner and A. Eden Eds.. Pages 1-22
- Dodig-Crnkovic G. and Müller V. (2009) A Dialogue Concerning Two World Systems: Info-Computational vs. Mechanistic. Book chapter in: *INFORMATION AND COMPUTATION*. World Scientific Publishing Co. Series in Information Studies. Editors: G Dodig-Crnkovic and M Burgin, 2011. <http://arxiv.org/abs/0910.5001>
- Dunbar R. (1998) *Grooming, Gossip, and the Evolution of Language*, Harvard Univ. Press
- Epstein, J. M. (2007). *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University.
- Gilbert N. and Troitzsch K. (2005) *Simulation for the Social Scientist*, Open University Press.
- Gilbert N: (2005) "Agent-based social simulation: dealing with complexity", <http://www.complexityscience.org/NoE/ABSS-dealing%20with%20complexity-1-1.pdf>
- Grimm V. and Railsback S. F. (2005) *Individual-based Modeling and Ecology*, Princeton University Press.

- Licklider, J.C.R. and Taylor R. W. (1968) "The computer as a communication device." *Science and Technology* (September), 20-41.
- Lim H. C., Stocker R., Larkin H. (2008) "Ethical Trust and Social Moral Norms Simulation: A Bio-inspired Agent-Based Modelling Approach. " In: 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, December 2008. pp. 245-251.
- Miller J. H. and Page, S. E. (2007) "Complex Adaptive Systems: An Introduction to Computational Models of Social Life", Princeton University Press: Princeton, NJ.
- Riedl J. (2011) "The Promise and Peril of Social Computing," *Computer*, vol.44, no.1, pp.93-95, <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5688159&isnumber=5688134>
- Schuler D. (1994) "Social Computing," *Comm. ACM*, vol. 37, no. 1, pp. 28–29.
- Wang F-Y., Carley K. M., Zeng D., and Mao W. (2007) "Social Computing: From Social Informatics to Social Intelligence. " *IEEE Intelligent Systems* 22, 2 (March 2007), 79-83. DOI=10.1109/MIS.2007.41 <http://dx.doi.org/10.1109/MIS.2007.41>.
- Zhang D., Guo B., Yu Z. (2011) "Social and Community Intelligence." *Computer*, Vol. 99, No. PrePrints. doi:10.1109/MC.2011.65.

## OBJECTS OF IDENTITY, IDENTITY OF OBJECTS

*For a Materialist Account of Online Behavior*

HAMID R. EKBIA  
*hekbia@indiana.edu*  
*School of Library and Information Science*  
*Indiana University Bloomington, IN. 47401*  
*U.S.A.*

AND

GUO ZHANG  
*guozhang@indiana.edu*  
*School of Library and Information Science*  
*Indiana University Bloomington, IN. 47401*  
*U.S.A.*

**Abstract.** Objects constitute significant elements of individual identity. Who we are has a lot to do with what we have and with what value we put on what we have. This point is easier to appreciate in the “off-line” physical world where objects with various symbolic or non-symbolic values populate our environment. How about the online world, which is seemingly devoid of objects — at least in a purely physicalist understanding of objecthood? What role, if any, do objects play in shaping online identities? We seek to address this question by following two lines of inquiry: post-structuralist accounts of quasi-objects and recent work in economic sociology on justification and mutual agreement. These inquiries lead to two key propositions: (i) Digital artifacts are quasi-objects, which *mediate* collective practices that seem to exert a strong force of desire in the specific circumstances of our times; and (ii) People operate within various regimes in which they *enact* information and objects through collective practices of situated social orders. Here we integrate and extend these two lines of inquiry in order to explore the question of online identity. Our key argument is that people’s identities are mediated through digital artifacts (personal websites, personal profiles, blogs, etc.) in a process in which the identities of the subject *and* the object are collectively and mutually enacted by the network of people who take interest in them.

## 1. Introduction

Objects constitute significant elements of individual identity. Who we are has a lot to do with what we have and with what value we put on what we have. This point is easier to appreciate in the “off-line” physical world where objects with various symbolic or non-symbolic values populate our environment. How about the online world, which is seemingly devoid of objects — at least in a purely physicalist understanding of objecthood? What role, if any, do objects play in shaping online identities?

We take this question seriously, and seek a materialist answer to it. We seek an account that can do justice to *things that matter*, that offer potentials and resistances, physically but also socially, historically, psychologically, and so on. Although this is admittedly a non-standard notion of materialism — modern philosophers often use physicalism and materialism interchangeably (Stoljar, 2009) — it is useful for our purposes in at least two ways. First, it allows us to consider the inherently material, not necessarily physical, aspects of the online world. Second, it opens a line of inquiry that situates digital artifacts in how they relate to existing social structures and in how they embody and anticipate the future through the socio-material practices that they allow or disallow. The first point is important because dominant discourses in information science, philosophy, and elsewhere tend to discount the underlying materiality (even physicality) of the “virtual” (e.g., Lévy, 1998). The second point matters because it allows us to see current online experiences from the historical perspective of modernity (Day and Ekbia, 2010).

## 2. Two Lines of Inquiry

Our study of the relationship between objects and identity in the online world follows two lines of inquiry. One is inspired by post-structuralist accounts of quasi-objects, the other by recent work in economic sociology on justification and mutual agreement.

Originating in the psychoanalytic notion of “part-objects,” Winicott’s notion of “transitional object,” and the Lacanian notion of *objet petit a* (object little-a), the notion of “quasi-object” later appears in discussions of intersubjectivity by Serres, of scientific theories and entities by Latour, and of technology and virtuality by Lévy. In Lacan’s (1991) psychoanalysis, *objet petit a* stands for an unattainable libidinal object of desire (e.g., the breast), which is imagined to be separable from the rest of the body, in the same fashion that an ornament can be detached from the body. As such, it both drives and limits the desire, and can be sought in the “other” traversing the order of the real and the imaginary, the mind and the body, the self and the other. In the age of the Internet, this raises the question of whether our common fascination and obsession with online depictions of our identity — digital variants of Lacan’s “mirror image” — may be a reassertion of specific (infantile?) desires. Answering this question in earnest requires empirical research on how identities are fluidly (de-, re-)constructed on the Web (Aboujaoude, 2011). However, the beginnings of an answer can be found in the writings of Michel Serres (1982) who seeks to explain identity and intersubjectivity from a materialist perspective. Famously characterizing the furet in a children’s game (a French game resembling hunt-the-slipper) as a quasi-object, Serres argues that the identity of the child who carries the furet changes as he becomes distinct from others by becoming “it”

(Serres, 1982). In so doing, the furet also connects the players and their positions, fixing and stabilizing the collective. The passage of the furet, in other words, allows the co-constitution of both (quasi-)objects and (quasi-)subjects (Day 2010).

Economic sociology, on the other hand, shows that subjects and objects are mutually qualified in different orders of worth. In their attempt to integrate economic and social values in a single analytic framework, for instance, Bolatanski and Thévenot (2006) have arrived at a set of principles that people resort to in order to justify their actions. These principles, which operate within different regimes of worth, are appealed to by individuals depending on the particular “world” (or polity) in which they inhabit in a given situation. “Persons and things offer one another mutual support. . . With the help of *objects*, which we shall define by their belonging to a specific world, people can succeed in establishing states of worth.” (Bolatanski and Thévenot, 2006: 131).

In previous work, these lines of thought have led us to two key propositions: (i) Digital artifacts are quasi-objects, which *mediate* collective practices that seem to exert a strong force of desire in the specific circumstances of our times (Ekbia, 2009a); and (ii) People operate within various *regimes of information* in which they enact information through collective practices of situated social orders (Ekbia, 2009b; Ekbia and Evans, 2009; Garfinkel, 2008). Here we integrate and extend these two lines of inquiry in order to explore the question of online identity. Our key argument is that people’s identities are *mediated* through digital artifacts (personal websites, personal profiles, blogs, etc.) in a process in which the identities of the subject *and* the object are collectively and mutually *enacted* by the network of people who take interest in them.

### 3. Online Behavior: Game and Identity

Take your personal profile on a social networking site, for instance. The profile *represents* you, but not in the sense that your photograph, for example, would represent you. By creating a profile, in a way you create a representation of yourself, your history, tastes, hobbies, friends, friends of friends, and so on. But on closer scrutiny this is *not* a representation, traditionally understood as a stand-in that has a resemblance relationship to you. Nor is the profile simply an active representation noncausally coupled to you in the way that most computer representations are believed to be coupled to their subject matter. The profile is an artifact that both mediates and traces your network of friends, hobbies, and history. As a complex event, not a representation, it constitutes a complex site for the actualization of such a network. Lastly, the profile participates in the embedding environment, taking you to unforeseen places, while being itself shoved around by others. In this manner, it acts like characters in a good novel who take on, we are told, a life of their own, dragging the author along with them (Bakhtin, 1984). In a serious way, the fate of the profile is in the hands of others who take interest in it and who build bridges between you and their profiles. In short, your identity is enacted in a collective process organized around your profile, in the same way that the identity of the child is shaped in carrying the furet. You become “it,” with the caveat that the nature of the “it” in an electronic medium enables a strongly malleable, transient, and unstable identity, providing enormous room for playfulness, fantasy, illusion, deception, self-deception, and so forth.

We want to explore these issues, especially in regards to computer games and how

an individual's "virtual" identity in a game may, or may not, interact with their identity in the non-game (off-line) world. With the growing potential of personalizing game characters (avatars) to represent individual features, this question has become increasingly meaningful and significant. For instance, in games for health, we can connect a Personal Health Record to a gaming platform so that, through proper data linkages to environmental signals, one's real-life behavior would affect the game — think of an avatar that becomes large, drunk, or ill depending on how you eat, drink, or behave. How would the change of the avatar influence your real-life identity? Is the avatar the equivalent of the furet? Or does it exert less/more influence?

## References

- Aboujaoude, E. (2011). *The Dangerous Powers of E-Personality*. New York: W.W. Norton & Company.
- Bakhtin, M.M. (1984). *The Problems of Dostoevsky's Poetics*. (C. Emerson, Trans.). University of Minnesota Press.
- Bolatanski and Thévenot Boltanski, L., &Thévenot, L. (2006 [1991]). *On Justification: Economies of Worth*. (C. Porter, Trans.). Princeton, NJ: Princeton University Press
- Day, R. E. (2010). Death of the User: Reconceptualizing subjects, objects, and their relations. *Journal of the American Society for Information Science and Technology*, 62(1)-78-88.
- Day, R. & Ekbia, H. (2010). Digital experiences. In *Kallinikos, J., Lanzara, G. F. and Nardi, B.* (Ed.). *The digital habitat — Rethinking experience and social practice. First Monday*, Volume 15, Number 6 - 7.
- Ekbia, H. (2009a). Digital artifacts as quasi-objects: Qualification, mediation, and materiality. *Journal of American Society for Information Science and Technology*, 60 (12), 2554-2566.
- Ekbia, H. (2009b). Regimes of information: A polity model. Paper presented at the 7<sup>th</sup> European Conference on Computing and Philosophy, Barcelona, Spain, July 1-4.
- Ekbia, H., & Evans, T. (2009). Regimes of information: Land use, management, and policy. *The Information Society*, 25, 328-343.
- Garfinkel, H. (2008). *Toward a sociological theory of information*. Boulder, CO: Paradigm.
- Lacan, J. (1991). The seminar of Jacques Lacan. In *Book II: The ego in Freud's theory and in the technique of psychoanalysis* (pp. 1954–1955). New York: W.W. Norton & Company.
- Lévy, P. (1998). *Becoming virtual: Reality in the digital age*. (R. Bonnono. Trans.) NewYork: Plenum.
- Serres, M. (1982). *The parasite*. (L. R. Schehr. Trans.). Baltimore: Johns Hopkins University Press.
- Stoljar, D. (2009). Physicalism. *Stanford Encyclopedia of Philosophy*. Retrieved March 24, 2010 from: <http://plato.stanford.edu/entries/physicalism/>

## THE CONSTRUCTION OF REALITY AND OF SOCIAL BEING IN THE INFORMATION AGE

LÁSZLÓ ROPOLYI

*Department of History and Philosophy of Science  
Eötvös University, 1518 Budapest, Pf. 32., Hungary  
ropolyi@caesar.elte.hu*

**Abstract.** In the information age representational (information, cognitive, cultural, communication) technologies instead of material ones become the dominant factor in the construction of social being. To conceptualize this shift, I suggest that Aristotle's dualistic ontological system (which distinguishes between actual and potential being) be complemented with a third form of being: virtuality. In the virtual form of being, actuality and potentiality are inseparably intertwined. Everything that is produced by representational technologies is a virtual being. Therefore, in the information age, social being, too, has a virtual character, as it is produced by representational technologies. Information itself is a product of representational technology; while it is also interpreted being. This process of interpretation takes place in human minds, and the process can be described as a "hermeneutical industry". The information society is inhabited by virtual beings, so it has a virtual and open characteristic.

### 1. Technology and Representation

Technology is a specific form or aspect of human agency, the realization of the human control over a technological situation.<sup>18</sup> Every element of the human world is created by technologies. Both human nature and the social being are the products of our technological activity, and their characteristics are determined by the specificities of the technology we use to produce them.<sup>19</sup>

All historical forms of human nature and of social being are constructed (and continuously re-constructed) or produced (and continuously re-produced) by historical versions of technology. Technology has an ontological *Janus face*: it produces both

---

<sup>18</sup> This definition of technology is on a higher level of abstraction than usual conceptualizations (cf. Feenberg, 1999).

<sup>19</sup> Social (or human) being, obviously, has an active role in the formation of any technology: given technological and social relations coexist and interrelate in a complex way, so that they mutually shape each other. My view on construction is closer to that of Marxism (Lukács 1978) than to those of phenomenology (Berger and Luckmann, 1966) and of radical constructivism (Glaserfeld, 2011).

“things” and “representations”. For thousands of years, people used material (agricultural or industrial) technologies where the material product was in the foreground, although the symbolic content was also present.

The last few decades have witnessed a significant technological change, in that “representations” have become dominant over the “thingly” products in the most important technologies of our age. On the one hand, new (*cognitive, communication, cultural, and information*) technologies have emerged; on the other hand, the representational or symbolic function of traditional technologies has become more significant. As a consequence, the most important characteristics of the social being are essentially transformed. The terms “post-industrial / knowledge / risk / information / network society” all refer to a type of society where representational technologies are the dominant factor in the (re)construction or the (re)production of human nature and of social being.

## 2. Virtuality and Openness in Information Technologies

The shift from material technologies to representational (information, cognitive, cultural, communication) technologies has important consequences for our notions of *reality*. The concept of *virtuality* has a central role in redefining reality. The term “virtuality” is relatively new, but a brief overview of the history of philosophy reveals that the fundamental components of virtuality have been extensively discussed (Ropolyi, 2001). The central concepts in this respect are presence, worldliness, and plurality. All three acquire their meaning from a certain relation between actuality and potentiality.

I suggest that the Aristotelian dualistic ontological system, which distinguishes between actual and potential being, be complemented with a third form of being: virtuality. In the virtual form of being, actuality and potentiality are inseparably intertwined. *Virtuality is potentiality considered together with its actualization*. Openness is actuality considered together with its possibilities. As compared to reality, virtuality is *reality with a measure*, a reality which has no absolute character, but which has a relative nature.

All beings produced by representational technologies are necessarily virtual. To illustrate how technologies produce virtual beings, let us consider information technologies. The characterization of information technology should be based on an understanding of the concept of information. Obviously, information is a product of a kind of representational technology, and thus it is virtual. In a hermeneutic approach, *information is “interpreted being”*. On this account, information technology is a “hermeneutical industry”, where the production is performed by interpretation in the minds of people. All the products of this “industry” are virtual beings. Consequently, social being in the information age is necessarily a virtual being. Information society is a society where the typical beings are virtual ones, and so the whole society has a virtual and open characteristic.

In a specific point of view *the Internet*, too, is a kind of information technology. It is an intentionally created and maintained artificial, virtual sphere which is based on networked computers and individual human interpretation praxes. The Internet is the medium (or sphere) of a new, virtual mode of human existence, basically independent

from, but built on, and coexisting with the former (natural and societal) spheres of existence, and created by the late-modern humans.

### **Acknowledgements**

This research was supported by the Hungarian Scientific Research Fund (OTKA) under the K79194 and K 84145 project numbers.

### **References**

- Berger, P. & Luckmann, T. (1966). *The Social Construction of Reality. A Treatise in the Sociology of Knowledge*. New York: Doubleday.
- Feenberg, A. (1999). *Questioning Technology*. London: Routledge.
- Glaserfeld, E. von (2011). <http://www.vonglasersfeld.com/> (March 2011).
- Lukács, G. (1978). *The Ontology of Social Being*. London: The Merlin Press.
- Ropolyi, L. (2001). Virtuality and plurality. In: A. Riegler, M. F. Peschl, K. Edlinger, G. Fleck and W. Feigl (Eds.), *Virtual Reality. Cognitive Foundations, Technological Issues & Philosophical Implications*. (pp. 167-187). Frankfurt am Main: Peter Lang.

## TRUST, KNOWLEDGE AND SOCIAL COMPUTING

### *Relating Philosophy of Computing and Epistemology*

JUDITH SIMON

*Institut Jean Nicod – Ecole Normale Supérieure  
29, rue d'Ulm  
F-75005 Paris - France*

**Abstract.** The main goal of my talk will be to link the discourse on trust in epistemology with the philosophical discourses on trust and ICT. I will argue that linking these two lines of research is needed to apprehend the notion of epistemic trust. Epistemic practices in science as well as in everyday life are characterized not only by their socialness, i.e. the fact that agents collaborate and rely on others in their attempts to know, they are also deeply pervaded by information technologies. In short, I claim that a) contemporary epistemic practices take place in increasingly complex, dynamic and entangled socio-technical epistemic systems consisting of multiple human and non-human agents, b) that trust is a crucial concept to understand these practices, and c) that information and communication technologies (ICT) play an important role in mediating and shaping trust relationship between different agents.

### **1. Trusting to Know**

In 1991, Hardwig asserts that “[f]or most epistemologists, it is not only that trust plays no role in knowing: trusting and knowing is deeply antithetical. We can not know by trusting in the opinions of others: we may have to trust those opinions when we do not know ((Hardwig 1991): 693). This argument rests on the assumption that in order to know, we have to be able to provide evidence, we have to justify our knowledge claims with our own cognitive resources and cannot know by simply trusting the testimony of others. Yet a closer look on epistemic practices in science as well as in everyday life shows that our knowledge depends deeply on trust in other people. Without trusting in what others have told us, we would neither know some of the most basic facts about ourselves, such as the date and place of our birth, nor could we have achieved the most advanced scientific knowledge. This is the central dilemma of testimony and epistemic trust in philosophy: while on the one hand it seems that almost everything we know depends on our trust in the testimony of others, the status of testimonial knowledge and the role of epistemic trust remain highly controversial. Yet things are even more complicated. Within contemporary epistemic practices trust is not only placed in other

humans, but also in technologies, processes, institutions and content. Indeed, information and communication technologies (ICT) play a special role for epistemic trust, because ICT is not only an entity that can be trusted itself, ICT also increasingly mediates and shapes trust relations between all other entities as well. Hence, to understand epistemic trust, the role of ICT cannot be ignored and epistemology has to take insights from other fields of research, most notably philosophy of computing and into account.

## 2. Trust and ICT

The special role of ICT for trust has been addressed under different labels such as online trust, digital trust or e-trust. While all terms refer to practices of trust that take place in a digital environment, the different labels are related to different research foci. Three of them should be distinguished:

1. ICT as an entity of trust itself (i.e. how human agents place trust in ICT as a technology)
2. ICT as a mediator of trust relationships between human agents as well as between humans agents and other entities (such as content)
3. Trust in multi-agent systems, i.e. trust relations amongst artificial agents as well as between human and artificial agents

First, ICT can be an entity that is trusted itself, i.e. trust into ICT can be considered as trust in a specific type of technology, hence as a special case of trust in technologies. Here analyses of whether one can rightfully talk about trust in technology in the first place (for instance (Nissenbaum 2001), or whether and to what extent we do or should place trust in technologies have been discussed ((Cheshire, Antin et al. 2010)).

Second, ICT mediates trust relations amongst and between humans and non-human entities to a profound extent. Even in the most basic form, if communication between two humans who know each other in person takes place via email, chat, social networking sites or even telephone, ICT mediates between truster and trustee (cf. (Ess 2010)). Epistemic trust placed in such technologies cannot be fully understood by referring to trust in technology or trust in persons only. Take the example of the online-encyclopedia Wikipedia. If one trusts content from Wikipedia, this practice of trust is neither trust in a technology proper (namely the wiki-software), nor is it trust in individual writers (which are often unknown), nor can this trust be fully explained by institutional trust in the Wikimedia Foundation. I have argued elsewhere, that trusting Wikipedia should rather be conceived as trust into a certain socio-technical epistemic system characterized by technological infrastructure, epistemic agents (i.e. the users of Wikipedia), and certain processes employed in creating epistemic content ((Simon 2010b)).

While Wikipedia ((de Laat 2010), (Tollefsen 2009), (Magnus 2009)) and Blogs ((Goldman 2008)) have attracted some interest within epistemology by now, other types of social software, such as recommender systems or social tagging systems have not yet received serious attention. Yet, in such types of social software that function primarily via aggregation, problems of trust are potentially even harder to tackle and the classical means provided by epistemological analyses on trust in testimony appear even less suited for understanding epistemic trust within such applications.

Finally, there is another type of e-trust, which is starting to receive attention within philosophy: trust in multi-agent systems. Two instances of trust are crucial with respect to trust in multi-agent-systems. First, there are the trust relations amongst artificial agents within multi-agent-systems. (e.g. (Taddeo 2010b)). Second, there are not only trust relations amongst artificial agents, but also between human and artificial agents, which are intrinsically more complex as (Grodzinsky, Miller et al. 2010) have noted.

In my talk I will specify in more detail, how these insights from the philosophy of computing could be made useful for an epistemology of trust.

## References

- Cheshire, C., Antin J. et al. (2010) General and Familiar Trust in Websites. *Knowledge, Technology & Policy* 23(3), 311-331.
- de Laat, P. (2010). How can contributors to open-source communities be trusted? On the assumption, inference, and substitution of trust. *Ethics and Information Technology* 12(4): 327-341.
- Ess, C. (2010). "Trust and New Communication Technologies: Vicious Circles, Virtuous Circles, Possible Futures. *Knowledge, Technology & Policy* 23(3): 287-305.
- Goldman, Alvin (2008). The Social Epistemology of Blogging. In: *Information Technology and Moral Philosophy*. J. v. d. Hoven and J. Weckert. New York, Cambridge University Press: 11-122.
- Grodzinsky, F., K. Miller, et al. (2010). "Developing artificial agents worthy of trust: —Would you buy a used car from this artificial agent?□." *Ethics and Information Technology*: 1-11.
- Magnus, P. D. (2009). On Trusting Wikipedia. *Episteme* 6(1): 74-90.
- Nissenbaum, H. (2001). Securing Trust Online: Wisdom or Oxymoron. *Boston University Law Review* 81(3): 635-664.
- Simon, J. (2010b). The entanglement of trust and knowledge on the Web. *Ethics and Information Technology* 12(4): 343-355.
- Taddeo, M. (2010b). Modelling Trust in Artificial Agents, a First Step toward the Analysis of e-Trust. *Minds and Machines* 20(2): 243-257.
- Tollefsen, D. P. (2009). Wikipedia and the Epistemology of Testimony. *Episteme* 6(1): 8-24.

## OPERATIONAL IMAGES

### *Agent-Based Computer Simulation and the Epistemic Impact of Dynamic Visualization*

SEBASTIAN VEHLKEN

*Leuphana University Lüneburg*

*ICAM Institute for Culture and Aesthetics of Digital Media*

*Scharnhorststrasse 1*

*21335 Lüneburg*

### **Abstract**

Computer simulations (CS) designate the current scientific condition. Inevitably, one has to distinguish crash tests from climate simulations, and one has to be aware of the differing problem dimensions posed by e.g. the simulation a quantum physical system by a classical physical system in comparison to those advanced by an agent-based simulation of a mass panic in a stadium. And without question, CS achieve diverse tasks and have quite dissimilar reputations in different scientific disciplines. But undeniably, CS brought with them a novel kind of *knowledge*, a modified set of *research problems*, and a transformed historical-philosophical *comprehension of science*. Thus, knowledge emerging in CS derives from the computer-based imitation of dynamic system behavior which penetrate everyday life in forms of ecological, medical, economical, or technical applications and decisions. Initially, novel scientific problems and research fields historically form where they would not have been tractable without the digital media of CS. And not least, the traditional concepts of theory and experiment are essentially modified, transforming the „mode-1« science (Gibbons, 1994) more and more into a „behavioral science of complex systems“ (Mahr, 2003). This transformation is based on an explicitly media-historical rupture marked by the digital mediality of CS. The digital media inherent in CS develop typical and intrinsic modes of operation and visualization in their application on analytically and experimentally intractable problem fields. Sebastian Vehlken’s presentation embarks on examining the “social computing” aspects of a particular kind of CS in a two-fold way. First, it will describe the specific (self-) organizational aspects of agent-based modeling and simulation (ABM), zeroing in on several pivotal examples of large-scale *social simulations*. These range from crowd control (e.g. *Massive Insight*) and logistics (e.g. *TransSims*) to epidemics (e.g. *PLAN-C* by NYU Bioinformatics Group) and large-scale models of the complex interactions of agents in whole societies (e.g. *Global Scale Agent Model* by *Brookings Institution*). It will discuss the notion, the epistemic function and the technological means of the bottom-up modeling paradigm of ABM, providing essential advantages over CS based on discrete events. Whilst the latter are required to define assumptions of the constituents of a system and their interdependencies from top down, ABM are decentralized and

function without a definition of the global system behavior. The system behavior emerges from the definition of simple and locally (on the level of the individual agents) implemented settings. As Borshchev and Filippov (2004) put it, ABM thus better »provides for construction of models in the absence of the knowledge about the global interdependencies: you may know nothing or very little about how things affect each other at the aggregate level, or what is the global sequence of operations, etc., but if you have some perception of how the individual participants of the process behave, you can construct the AB model and then obtain the global behavior.« The bottom-up performance of ABM induces a synthetic problem approach by converging to adequate and context-dependent solutions in a process of a systematic comparison and evaluation of different simulation runs and scenarios. Thereby ABM leapfrogs fixed object or context allocations in an exemplarily interdisciplinary manner. The media history of research in social collectives reveals a reciprocal ›socialization‹ and ›biologization‹ of computer science and a likewise computerization of the social sciences when it comes to the development of adequate ABM models for describing collective behaviors in space and time. The development of Animation Effects in CGI is distinctly interconnected with biological and sociological computer models of collective dynamics, and vice versa. Second, it will consider the importance of *digital visualizations* for scientific research with ABM. The adherent types of Computer Graphical Imagery (CGI) exemplarily raise questions not only about the status of animated, 3-dimensional and dynamic digital images as interfaces for the refinement of societal “computer experiments” and the “intuitive” handling of the ABM by researchers. One must also ask about their state as ‘visual evidence’ and ‘representation’ for phenomena and processes in social dynamics which would remain intractable without these digital ‘time-based images’. Not least, the technological conditions resulting of the multiple filtering-, smoothing-, or thresholding procedures involved in providing ‘visual validation’ have to be accounted for. These aspects have to be further investigated on the basis of a media-technologically informed theory of *operational images*, linking the modes of visualization of ABM with their programmed data base in the ABM software. And since the development of certain Animation Effects in the CGI industry is historically distinctly interconnected with biological and sociological computer models of collective dynamics, and vice versa, the hard-, wet- and software foundations of ABM can be short-circuited with applicable modes of CGI generation: both operate in a highly distributed manner of ›socially‹ interacting and ›locally‹ defined agents. Hence, the presentation investigates the specific epistemical and technological rupture marked by CS on the basis of ABM in *social simulations*. The respective applications facilitate a mode of visualization by (synthetic and therefore operational) images which address the inconceivable *representation* of complex social dynamics by generating visual *presentations*: Only the observation of modeled processes in the runtime of ABM enables the evaluation and manipulation of critical factors and variables and the ensuing re-run of the simulation. And this results in a type of dynamical “data images” (see Adelman et al., 2009, Schubbach, 2007) yet to be further investigated. It provokes a type of *operational images* with a highly socio-political dimension – images

which depend on and which foster social decision-making in (time-) critical environments.

## References

- Adelmann, R., Frercks, J., Heßler, M. & Henning, J. (Eds.)(2009). *Datenbilder. Zur digitalen Bildpraxis in den Naturwissenschaften*, Bielefeld 2009.
- Borshchev, A. & Filippov, A. (2004). From System Dynamics and Discrete Event to Practical Agent Based Modeling: Reasons, Techniques, Tools. In: *The 22nd International Conference of the System Dynamics Society*. Oxford.
- Mahr, B. (2003). Modellieren. Beobachtungen und Gedanken zur Geschichte des Modellbegriffs. In: H. Bredekamp and S. Krämer (Eds.), *Bild Schrift Zahl* (pp. 59-86). Munich: Fink.
- Schubbach, A. (2007). ...A Display (Not a Representation)... *Navigationen. Zeitschrift für Medien- und Kulturwissenschaft. Display II – digital* 7(2) (2007, 13–27).

## Social Computation as a Discovery Model for the Social Sciences

AZIZ F. ZAMBAK  
*Department of Philosophy*  
*Yeditepe University, Istanbul*

**Abstract.** Social simulation is a growing field that proposes a computational approach to the social sciences. Simulation provides a powerful alternative for the novel understanding of the epistemology, ontology, and taxonomy of the social phenomenon, structure and process. Social simulation can be an *intellectual resource* and *experimental field* for developing a novel notion of “social phenomenon” within which various forms of human action can be represented. Social simulation may be used to examine not just the current situation in a society, but also possible social situations. Classical models that only use natural language is inadequate for the comprehension of dynamic and complex systems in the social sciences. Pure mathematical and/or statistical models are intractable. Simulation may offer to overcome the limitations of classical models in the social sciences. In this paper, we will propose five general principles that should be take into consideration in social simulation: 1- Agent-Based Models: We describe agency as an essential criterion for social simulation. 2- Game Theory: Game theory is a study that can provide some formal epistemological data for understanding the rationalization process of individuals. From the social simulation point of view, discovery is an agentive-informational-system and we consider this system as a set of complex principles that should be *rationalized* by simplification, approximation, optimization, and generalization. 3- Control Systems: In order to understand the autopoietic, dynamic and complex structure of social systems, we should develop an organismic conception of society in which control mechanisms have an essential role for the social models and simulation. 4- Tools: In social simulation, a stylized-computational-language should be built in which the data on social structure are coded and represented in the computer simulation. 5- Ontology: Emergence is one of the essential concepts in the ontology of social sciences in which certain theories try to explain the macrolevel phenomena in terms of the behavior of microlevel actors.

Social simulation is a growing field that proposes a computational approach to the social sciences.<sup>20</sup> Simulation provides a powerful alternative for the novel understanding of the

---

<sup>20</sup> Gilbert and Troitzsch (2005: 5) explains the main reason behind the developing interest on social simulation as follows: “The major reason for social scientists becoming increasingly interested in computer simulation, however, is its potential to assist *discovery* and *formalization*. Social scientists can build vey simple models that focus on some small aspects of the social world and discover the consequences of their theories in the ‘artificial society’ that they have built. In order to do this, they need to take theories that have conventionally been expressed in textual form and formalize them into a specification which can be

epistemology, ontology, and taxonomy of the social phenomenon, structure and process. Social simulation can be an *intellectual resource* and *experimental field* for developing a novel notion of “social phenomenon” within which various forms of human action can be represented. Social simulation may be used to examine not just the current situation in a society, but also possible social situations. Classical models that only use natural language is inadequate for the comprehension of dynamic and complex systems in the social sciences. Pure mathematical and/or statistical models are intractable. Simulation may offer to overcome the limitations of classical models in the social sciences.

In this paper, we will propose five general principles that should be take into consideration in social simulation.

### **1- Agent-Based Models:**

Agency must be the central notion in social simulation since the cognition of social reality originates from agentive actions. We claim that agency is the ontological and epistemological constituent of social reality. It is characterized by agentive activity. Agency must be the essential criterion for the success of social simulation. Social simulation must consider the social phenomena as a form of action of a dynamic-representational system, developed during *interaction* within the environment. Equating properties of the social phenomena with properties of its elements [individuals] is the basic mistake. Social structure cannot be a subject of a special examination of the group of individuals. Behavior and agentive actions cannot be found in the specific groups of individuals, but in the whole agent-environment-interaction system. The discovery of social phenomena in social simulation does mean a new kind of action of the highly dynamic-representational system capable of making inferences from its structure and process in order to achieve new results of action and form novel systems directed towards the future. Therefore, in social simulation, discovery is not a mystical emergent property of social phenomena, but a form of agentive action necessarily following from the development of a dynamic-representational system.

### **2- Game Theory:**

Game theory is a study that can provide some formal epistemological data for understanding the rationalization process of individuals. From the social simulation point of view, discovery is an agentive-informational-system and we consider this system as a set of complex principles that should be *rationalized* by simplification, approximation, optimization, and generalization. In social simulation, this type of *rationalization* should depend on idealization. Idealization transforms the environmental data into ideal-agentive-rational-information. However, idealization should not be seen as abstraction.<sup>21</sup> We consider the idealized information as one of the basic capabilities of social simulation, providing the preconditions for the adaptive behavior of agency in a very

---

programmed into a computer. The process of formalization, which involves being precise about what the theory means and making sure that it is complete and coherent, is very valuable discipline in the social sciences to that of mathematics in the physical sciences.”

<sup>21</sup> As Nowak (2000: 116) states, “idealization is not abstraction. Roughly, abstraction consists in a passage from properties *AB* to *A*, idealization consists in a passage from *AB* to *A-B*.”

complex environment. In the adaptiveness of agency, the information of environmental structure and organization may be grasped *rationally*, for the *rationality* lies in the agentive attitude towards environmental structure and organization, not in the essence of environment itself. Therefore, there is not a hidden essence in the environmental structure and organization that should be represented in a computational and representational manner for the *rational* behavior of an agent. In social simulation, our aim is to understand how properties of *rationalized* agency are related to the behavioral action that is performed under complex environmental/social situations. This type of understanding requires idealization, as idealization can be seen as a method of constructing informational structures in which data gained from the environment/society can serve the goal of forming special types of *rationalized* agentive interactions. Idealization, in social simulation, leads an agent to a successful informational approximation. Idealization is a type of theorizing that includes specification, approximation and optimization about certain sets of agentive and social systems. The presentation will include analysis of two game theoretical models for social simulation.

### **3- Control Systems:**

Social systems should be considered as self-organizing, non-linear, dynamic, and complex phenomena. From the computational or representational point of view, dynamic and complex systems are difficult to study because most cannot be represented in simplified and hierarchical models. In order to understand the autopoietic, dynamic and complex structure of social systems, we should develop an organismic conception of society in which control mechanisms have an essential role for the social models and simulations. There are several conditions for choosing the appropriate strategy for the control mechanism of an agent such as the availability of data for the performance of an agent, comparing stable and dynamic parameters of the environment, and the access to explicit data about plans, goals, and the current state of affairs. For building computer simulation for an agentive system, it is very important not to restrict an agent to follow only one predetermined set of rules but to give it the opportunity to choose and shift different sets of rules according to its situation. This can be done by a proper control mechanism which can find a balance between stability and flexibility of information in a complex environment. In this section, we will also examine the *Project Cybersyn* as a control mechanism example for the social simulation.

### **4-Tools:**

In the presentation, we will briefly explain what should be the logic of computer programs in social simulation. In addition, we will claim that, in social simulation, a stylized-computational-language should be built in which the data on social structure are coded and represented in the computer simulation. The general concepts of this stylized-computational-language will be briefly introduced in the presentation. Some of these concepts are empirical protocols, nodes, links, data processing, boundaries, taxonomy, observation period, randomization of parameters, outcome validity, process validity, and internal validity.

## **5- Ontology**

Emergence is one of the essential concepts in the ontology of social sciences in which certain theories try to explain the macrolevel phenomena in terms of the behavior of microlevel actors. In this part, we will show that how a reflexive model in social simulation can build an emergent model of the relation between the individual and the society.

## **References**

- Gilbert, Nigel and Troitzsch, Klaus G. (2005). *Simulation for the Social Scientist*, Buckingham : Open University Press.
- Nowak, Leszek (2000). The Idealization Approach to Science: A New Survey. *Poznań Studies in Philosophy of Science and the Humanities*, 69, 109-184.

# **Track VIII: IT, Culture and Globalization**

## The Revival of National and Cultural Identity through Social Media

RYOKO ASAI

*Uppsala University, Dept. of IT-HCI*

*Box 337, 751 05 Uppsala, Sweden*

*and*

*Nihon University, College of Industrial Technology*

*Narashinoshi-Izumicho1-2-1, Chiba, Japan*

Iordanis Kavathatzopoulos

*Uppsala University, Dept. of IT-HCI*

*Box 337, 751 05 Uppsala, Sweden*

AND

Mikael Laaksoharju

*Uppsala University, Dept. of IT-HCI*

*Box 337, 751 05 Uppsala, Sweden*

**Abstract.** Social media has played an important role as hub for information in political change. It can contribute to the development of psychological and social preconditions for dialog and democracy.

Information communication technology (ICT) made it possible for people to communicate beyond national borders. In particular, social media play an important role in making a place where people communicate each other, for example Facebook, MySpace, YouTube and so on. In other words, under these circumstances, social media function as the third place (Oldenburg, 1999). People have two essential and indispensable places in their lives: one is home and another is working place. Further to those places, people have one more place where they could have relationships with others informally in public (what Oldenburg called “informal public life”). And the third place contributes not only to unite people in communities but also to know how they contribute in various problems and crises there. Therefore the third place would nurture a relationship with others and mutual trust under the unrestricted access condition, and also it would be open for discussion and ground for democracy (Oldenburg, 1999). In this context, social media can provide the third place to users in some cases.

Social contexts of communication are defined by geographic, organizational and situational variables, and those variables influence the contents of communication among people (Sproull & Kiesler, 1986). And, in order to discern social context cues, communicators observe static cues (physical setting, location etc.) and dynamic cues (non-verbal behavior like gesture or facial expression) in communicating with others. Communicators' behavior is determined based on social context cues and they can adjust

their behavior depending on situations through the process of interaction between them. However, in online communication, it is more difficult for communicators to perceive static and/or dynamic elements compared to face-to-face communication. Because in many cases social media limit the number of characters and the amount of data that they can post while making it possible for users to communicate regardless of physical distance, national boundaries and time difference. On the other hand, participation is seen as the key element in the recent trend toward democratization and in real numerous users send and receive a huge amount of information via social media to cultivate a relationship with others and strengthen mutual exchange beyond borders. In general, it is recognized that social media advance participation through exchanging information with minimal social context cues.

Tunisian people shared information on what happened in the country and when and where anti-government protests were held, by social media such as Facebook and twitter. In other words, social media seemed to support political change in Tunisia. Behind it, the number of the internet users is 3.6 million, which is 34% of the population total, and there are 1.6 million users of Facebook roughly equivalent to 16% of the population (Internet World Stats, 2010). Tunisian government had blocked particular websites. Facebook was one of the few social media free to access. Under these circumstances, for the people living abroad, Facebook functioned as primary source of information to have direct access to daily events in Tunisia.

Under these restrictive access conditions, social media like Facebook provides users with opportunities to communicate with others and also to state their opinion, in order to overcome constraint and the old regime. In this context, social media serve as the third place and users develop solidarity and reinforce identity through online communication. As is obvious from the statistical data on the internet users mentioned above, it is estimated that the number of in-country users of Facebook are fewer than the number of users living abroad. Many users followed with what was going on in Tunisia showing in-country users that they were all caring about political change. And this phenomenon is recognized as a kind of participation to collective movement through social media regardless of physical distance or time difference.

However, communication through social media has some problems. At first, exchanged information via social media is minimized social context cues under severe restricted conditions, due to sending information certainly and rationally. Therefore information tends to be extreme and there is a risk of group polarization. Second, in social media, information receivers gather fragmented information based on personal experience and make it plausible to understand easier as their own experience or to relive the experiences of its senders. And, through this process, users develop a sense of solidarity and share expectation as well as norms organizing them as one community. Therefore social norms accrete influence on users in particular communities and advance self-stereotyping among them as solidarity and social identity are enhanced. This situation is fraught with social risk of exclusion of others. Some people call Tunisian political change as “Facebook revolution” or “twitter revolution” on the internet. Are these diminutives really pertinent? Indeed, social media has played the important role as “hub for information” and the third place in political change. However, social media has to contribute to the development of skills for dialog in order to achieve a really democratic society (Asai & Kavathatzopoulos, 2010; Kavathatzopoulos, 2010, 2007).

## References

- Asai, R. and Kavathatzopoulos, I. (2010). Diversity in the construction of organization value. *Proceedings of EBEN Annual Conference 2010 "Which values for which organizations"*. Trento, Italy: University of Trento.
- Internet World Stats (2010). *Tunisia: Internet usage and marketing report*. Available online: <http://www.internetworldstats.com/af/tn.htm> (accessed February 7, 2011).
- Kavathatzopoulos, I. (2007). Information Technology as a tool for democratic skills. In A. Lionarakis (Ed.), *Forms of democracy in education: Open access and distance education* (pp. 155-162). Athens: Propobos.
- Kavathatzopoulos, I. (2010). Information technology, democratic societies and competitive markets. *Proceedings of the 3rd International Seminar on Information Law "An information law for the 21st century"*. Corfu, Greece: Ionian University.
- Kiesler, S. and Sproull, L. S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science*, 32(11), 1492-1512.
- Oldenburg, R. (1999). *The great good place*. Cambridge: Da Capo Press.

## WIKILEAKS AND ETHICS OF WHISTLE BLOWING

Patrick Backhaus

*School of Innovation, Design and Engineering, Mälardalen University,  
Sweden [pbs10002@student.mdh.se](mailto:pbs10002@student.mdh.se)*

and

*Paderborn University, Germany [bpatrick@campus.uni-paderborn.de](mailto:bpatrick@campus.uni-paderborn.de)*

AND

Gordana Dodig Crnkovic

*School of Innovation, Design and Engineering, Mälardalen University,  
Sweden [gordana.dodig-crnkovic@mdh.se](mailto:gordana.dodig-crnkovic@mdh.se)*

### 1. Extended Abstract

In a time in which the Internet pervades everyday life and information published is readable all over the world, it becomes very important to deal with ethical problems related to whistle blowing via the Internet. Although there are basic concepts like anonymity, privacy and freedom of speech, for every new kind of phenomenon we have to discuss its ethical aspects (Kizza, 2010)( Nadler and Schulman, 2006). A current example is the platform WikiLeaks which publishes a vast amount of secret documents. To evaluate ethics of WikiLeaks (Hanson and Ceppos, 2006)(WikiLeaks About), we will apply the following ethical approaches:

*The Utilitarian Approach*, focusing on the consequences that the publications of WikiLeaks have on the well-being of all parties that are affected directly or indirectly, so there are two sides to consider:

- On the one hand, the uncovering of misconduct and the increased transparency of the government are of such importance that the publications benefit society as a whole. So it alleviates the opinion making and leads to a greater understanding of governmental work.
- On the other hand the publications may threaten the national security and so harm society. They lead to a society with decreased integrity which may eventually result in less communication, more technical restrictions and so in less freedom.

To achieve a balance between both sides a potential approach could be that WikiLeaks reduces their amount of published data and classify the data more in detail. Further they could contact the company or government concerned before the publication, so that this party itself could acknowledge the misconduct.

*The Virtue Ethics Approach*, focusing on attitudes that develop our human potentials such as e.g. honesty, courage, faithfulness, trustworthiness and integrity. It is easy to see that WikiLeaks disregards these virtues in many different contexts. They are accused for putting people's lives at risk, publishing stolen data and degrading loyalty, privacy and integrity of data. The only virtue they undoubtedly represent is transparency which is not considered classical ethical virtue, but may be seen as an element of democracy. So WikiLeaks must ensure that the increased transparency gained by the publication is much more worth than all other aspects which will only be the case at severe misconduct by the concerned party that is made public as no other way of corrective action was available.

*The Information Ethics Approach*: From the point of view of Information Ethics, we can study how information is revealed/communicated in the networks of agents. Within approach we can ask questions such as: what is the function of "information hiding" and "encapsulation" such as found in Object Oriented Programming and any hierarchical organization? What would be the behavior of a society in which every agent would be connected with every other agent and share any information they have?

Interesting to observe is the *global character* of WikiLeaks, in a world regulated on the base of nations, which seem to act in a grey zone since the legal situation is unclear and different governments are still searching for a crime Julian Assange can be charged for.

In reality the issue of WikiLeaks (Kintzinger and Zepelin, 2010) (Greenberg, 2010) implies much more than an ethical discussion about whistle blowing and leaking, integrity and freedom of speech. WikiLeaks have become a symbol of a deep change in the publicity of information in the digital age, at least with the present-day technology. It has generated the greatest confrontation between the established order and the advocacy of the culture of the totally open Internet.

We are at the moment a part of the world where it is difficult to control and keep information secret and safe from eavesdropping and unauthorized use. Some of the relevant questions are: Has the institution of legal secret, business secret, military or organizational secret become obsolete? If yes, why? If no, how to protect information which should be protected? Who and how decides which information is worth making public and which is not? According to Assange (Bieber, 2010) (Fallows, 2010) personal integrity must be protected. Why not institutional integrity?

If leaking is a good democratic mechanism shall we not have leaks of WikiLeaks as well? And so on...a chain, or a loop of leaks? In a totally transparent world, how would information overload be managed? Shall we give up all trust? Or, equally important: Whom shall we trust?

Perhaps problems with information protection will lead us to a society where conversations are reduced to minimum and information less accessible as it has become obvious that anything can be made public. In the end, the result would be not an increase, but a decrease of freedom.

## References

- Bieber C. (2010) "Die Ethik des Lecks", 11.08.2010. der Freitag  
<http://www.freitag.de/kultur/1032-die-ethik-des-lecks>

- Fallows J. (2010) "More on Mullen, Twitter, and the Ethics of WikiLeaks", July 2010. <http://www.theatlantic.com/politics/archive/2010/07/more-on-mullen-twitter-and-the-ethics-of-wikileaks/60705>
- Greenberg A. (2010) An Interview With WikiLeaks' Julian Assange, Nov. 29 2010. Forbes. <http://blogs.forbes.com/andygreenberg/2010/11/29/an-interview-with-wikileaks-julian-assange>
- Hanson K. and Ceppos J. (2006) "The Ethics of Leaks," <http://www.scu.edu/ethics/publications/ethicalperspectives/leaks.html>
- Kintzinger A. and Zepelin J. (2010) "Stärkt Wikileaks die Freiheit?", 02.12.2010. Financial Times Deutschland <http://www.ftd.de/it-medien/medien-internet/pro-und-kontra-staerkt-wikileaks-die-freiheit/50200724.html>
- Kizza J. M. (2010) "Cyberspace, Cyberethics, and Social Networking," in Ethical and Social Issues in the Information Age. London: Springer London, ch. 11, pp. 221–246. [http://dx.doi.org/10.1007/978-1-84996-038-0\\_11](http://dx.doi.org/10.1007/978-1-84996-038-0_11)
- Nadler J. and Schulman M. (2006) "Whistle Blowing in the Public Sector," November 2006. [http://www.scu.edu/ethics/practicing/focusareas/Government\\_ethics/introduction/whistleblowing.html](http://www.scu.edu/ethics/practicing/focusareas/Government_ethics/introduction/whistleblowing.html)
- WikiLeaks About, [Online]. <http://wikileaks.de/about.html>

All links accessed on 2011 04 25

## INTERPRETING CODES OF ETHICS IN GLOBAL SOFTWARE ENGINEERING

### *Extended Abstract*

THIJMEN DE GOOIJER  
*Mälardalen University*  
*Högskoleplan 1, Västerås, Sweden*

**Abstract.** In global software engineering (GSE) groups of people from all over the world collaborate on the development of one system. For example, it is common for Western companies to send development work to Asia or Eastern Europe. Within these collaborations the differences between cultures and the problems these differences create, are plentiful. Because we expect that computing professional organizations codes of ethics are insufficiently adapted to GSE, we investigate the culture-relative interpretations of codes of ethics and the guidance they provide for global teams and collaboration. We analyze the codes of ethics of the ACM (US), CSI (India), IPSJ (Japan), HKCS (Hong Kong) and EI (Ireland). We look whether the codes explicitly address ethical dilemmas caused by global interactions, and investigate the ethical guidance provided by the codes. For the latter we apply them to three case questions that one could raise in a GSE setting. Our work differs from that of others in that it examines the practical applicability of codes of ethics instead of their contents and that our goal is not to study different culture-relative interpretations of just one problem. During our analysis we did not find imperatives that directly hinder global interaction, but unfortunately we were also unable to find any that sufficiently address this topic. Only one of the studied codes asks to consider cultural differences. While answering the case questions using the imperatives from the aforementioned codes, the cultural perspectives needed to interpret the words become clear, and we learn that little attention is given to the problems associated with global collaboration. We conclude that all studied codes would benefit from more explicit guidelines for those professionals that work in GSE.

### **1. Introduction**

Despite the globalization of the software engineering profession, most computing professional organizations are active in a limited number of countries and have their own code of ethics (CoE) or code of conduct (CoC). These codes are thus national in scope (Wheeler, 2003). According to a 1996 study as much as 78% of IS professionals use these codes in their ethical decisions (Joyce et al., 2003). At the same time, ethical

reactions and attitudes are influenced by culture and national origin (Christie et al., 2003; Nyaw & Ng, 1994)□. As a result ethical decision making is a complex endeavor in the current global IS practice (Wheeler, 2003). We expect that the codes have not kept up with the globalization of the profession.

To explore the possible difficulties computing professionals may encounter during their ethical decision making in global software engineering (GSE), we analyze the codes of ethics of five professional organizations and apply their codes to three case studies. We characterize our study by the following research questions.

- Do the studied codes specify culture-relative imperatives that could hinder or support global software engineering?
- Do the studied codes provide adequate ethical guidance for IT professionals in global interactions?

## **2. Related Work**

To our knowledge no studies exist that take a similar, practical approach to identify problems for global software engineers in computing professional CoE. Earlier work does compare codes (Oz, 1993), even in international settings (Joyce et al., 2003; Wheeler, 2003) and is discussed below. Work that combines codes of ethics with cultural influences can be found for example in (Arnold et al., 2007), which studies the views of western European accountants on actions prescribed by CoC based on their country of origin. It is found that these views differ significantly.

Case studies exist which review the ethical stance of different cultures on specific issues, for example, software piracy (Swinyard, Rinne, & Kau, 1990), but these studies either do not include CoE of computing professional organizations or do not have the goal to study their usefulness in decision making. Specific in another way are the case studies in (Anderson et al., 1993), which focus only on the ACM code.

### **2.1. COMPARING CODES**

Oz reviews four codes of US computing professional organizations finding flaws, moral dilemmas, and points for improvement (Oz, 1993). We differ from (Oz, 1993) in that we do not limit our study to US codes.

In their study comparing 27 international CoE Joyce et al. found only eight themes that were common to more than 50% of the CoE (Joyce et al., 2003). Compared to the work by Joyce et al. our work aims to identify problems encountered during ethical decision making in a GSE context, while their work focusses on the content of the codes.

Wheeler (2003) compares the codes of the ACM, the British Computer Society (BCS) and the Australian Computer Society (ACS) to find differences and similarities. Our work differs from (Wheeler, 2003) in that we put more emphasis on how codes are used in a global setting and the selected codes.

### **2.2. A GLOBAL CODE**

Some voices suggest to unite everyone by one global code of ethics (Payne & Landry, 2006; Wheeler, 2003). Davison on the contrary does not believe it is possible to establish a global code due to differences between nations and cultures (Davison, 2000).

His concerns are supported by the difficulties IFIP experienced in the 90s when it attempted to establish a consensus document to serve as a base for the development of codes by member bodies (Joyce et al., 2003).

We consider the views of Brey (2007) and Wong (2009) more balanced. They both acknowledge that a universal ethic would be ideal, but respect that in practice this can only be implemented as an extension of the local moral systems (Brey, 2007) and that we should avoid to force 'our' ethics onto another culture (Wong, 2009).

### **3. Selection of CoE**

In our study we compare five CoE, those of: the Association for Computing Machinery (ACM, 1992), Computer Society of India (CSI, 2010), Hong Kong Computer Society (HKCS, 2010), Information Processing Society of Japan (IPSJ, 1996), and Engineers Ireland (EI, 2009). Only five codes were selected to limit the study to a manageable size. The codes are chosen based on the role of their organization's home country in GSE, as well as variation in culture. The full paper provides more rationale for the selection.

### **4. Static Code Analysis**

In this Section we answer our first research question. To do so we informally compare the content of the five codes. Our assumption is that if an imperative is culture-relative it will not appear in all codes. Note that this does not capture culture-relative interpretations of imperatives. It is to capture interpretation problems that we include the case studies in Section 5. Comparing the CoE we find that only one of them asks to consider cultural differences, but we find no imperatives that directly (by formulation) impede inter-cultural collaboration. A number is culturally bound, and we expect all will be interpreted differently even when imperatives match.

### **5. Employing The Codes**

In this Section we apply the five selected CoE to three case studies. In this way we hope to discover whether the studied codes provide adequate ethical guidance for IT professionals in global interactions. Below we formulate our case studies as three questions that one might ask him-/herself in a GSE project.

- Developing a medical system for deployment in several countries across the globe, should I be aware of all legal requirements?
- How do I design my system so that it respects the expected level of privacy?
- May I say 'yes' to an assignment I receive from a German customer when I am uncertain that I can complete it?

## 6. Concluding Remarks

While studying the CoE we found only a couple of imperatives that could hinder GSE collaboration. However, none of the codes seem to be written with global collaboration in mind. And only the IPSJ CoE explicitly mentions the problem of cultural differences. Further, the case studies show that decisions on ethical dilemmas will often depend on the interpretation by professionals or the implicit stance of the code. We feel that the CoE should provide more guidance to deal with the complexity of ethical decisions in a GSE setting. Our primary recommendation for computing professional organizations is to revise their CoE to reflect the advance of GSE. Future work could examine how this may best be achieved within each culture.

## Acknowledgements

A warm thanks to my professor Gordana Dodig-Crnkovic for encouraging me to submit this work to IACAP and her useful comments.

## References

- ACM. (1992). ACM Code of Ethics and Professional Conduct. Retrieved December 2010, from <http://www.acm.org/about/code-of-ethics>.
- Anderson, R. E., Johnson, D. G., Gotterbarn, D., & Perrolle, J. (1993). Using the new ACM code of ethics in decision making. *Commun. ACM*, 36(2), 98-107. New York, NY, USA: ACM. doi: <http://doi.acm.org/10.1145/151220.151231>.
- Arnold, D., Bernardi, R., Neidermeyer, P., & Schmee, J. (2007). The Effect of Country and Culture on Perceptions of Appropriate Ethical Actions Prescribed by Codes of Conduct: A Western European Perspective among Accountants. *Journal of Business Ethics*, 70(4), 327-340. Springer Netherlands. Retrieved from <http://dx.doi.org/10.1007/s10551-006-9113-6>.
- Brey, P. (2007). Is Information Ethics Culture-Relative? *International Journal of Technology and Human Interaction*, 3(3), 12-24.
- Christie, P. M. J., Kwon, I.-W. G., Stoeberl, P. A., & Baumhart, R. (2003). A Cross-Cultural Comparison of Ethical Attitudes of Business Managers: India Korea and the United States. *Journal of Business Ethics*, 46(3), 263-287. Springer Netherlands. Retrieved from <http://dx.doi.org/10.1023/A:1025501426590>.
- CSI. (2010). Computer Society of India - Code of Ethics. Retrieved December 2010, from <http://www.csi-india.org/web/csi/code-of-ethics>.
- Davison, R. M. (2000). Professional ethics in information systems: a personal perspective. *Commun. AIS*, 3(2es). Atlanta, GA, USA: Association for Information Systems. Retrieved from <http://portal.acm.org/citation.cfm?id=374504.374510>.
- EI. (2009). Engineers Ireland - Code of Ethics. Retrieved December 2010, from <http://www.engineersireland.ie/about-us/governance/code-of-ethics-and-by-laws/>.
- HKCS. (2010). Hong Kong Computer Society - Code of Ethics and Professional Conduct. Retrieved December 2010, from [http://www.hkcs.org.hk/en\\_hk/intro/coe.asp](http://www.hkcs.org.hk/en_hk/intro/coe.asp).
- IPSJ. (1996). Code of Ethics of the Information Processing Society of Japan. Retrieved December 2010, from [http://www.ipsj.or.jp/english/somu/ipsjcode/ipsjcode\\_e.html](http://www.ipsj.or.jp/english/somu/ipsjcode/ipsjcode_e.html).

- Joyce, D., Blackshaw, B., King, C., & Muller, L. (2003). Codes of Conduct for Computing Professionals: an International Comparison. In S. Mann & A. Williamson (Eds.), *Proceedings of the 16th Annual NACCQ, Palmerston North, New Zealand* (pp. 71-78).
- Nyaw, M.-K., & Ng, I. (1994). A comparative analysis of ethical beliefs: A four country study. *Journal of Business Ethics, 13*(7), 543-555. Springer Netherlands. Retrieved from <http://dx.doi.org/10.1007/BF00881299>.
- Oz, E. (1993). Ethical standards for computer professionals: A comparative analysis of four major codes. *Journal of Business Ethics, 12*(9), 709-726. Springer Netherlands. Retrieved from <http://dx.doi.org/10.1007/BF00881385>.
- Payne, D., & Landry, B. J. L. (2006). A uniform code of ethics: business and IT professional ethics. *Commun. ACM, 49*(11), 81-84. New York, NY, USA: ACM. doi: <http://doi.acm.org/10.1145/1167838.1167841>.
- Swinyard, W. R., Rinne, H., & Kau, A. K. (1990). The morality of software piracy: A cross-cultural analysis. *Journal of Business Ethics, 9*(8), 655-664. Springer Netherlands. Retrieved from <http://dx.doi.org/10.1007/BF00383392>.
- Wheeler, S. (2003). Comparing Three IS Codes of Ethics - ACM, ACS and BCS. *PACIS 2003 Proceedings* (p. Paper 107).
- Wong, P.-H. (2009). What should we share?: understanding the aim of Intercultural Information Ethics. *SIGCAS Comput. Soc., 39*(3), 50-58. New York, NY, USA: ACM. doi: <http://doi.acm.org/10.1145/1713066.1713070>.

## **INFORMATION TECHNOLOGY, GLOBALIZATION AND INTELLECTUAL PROPERTY RIGHTS**

SORAJ HONGLADAROM

*Department of Philosophy*

*Faculty of Arts, Chulalongkorn University*

The main concern of this paper centers around the issues arising from the use of intellectual property rights (IPRs) as a tool of globalization, and how creations of information technology are usually protected through the IPR regime as well as how the technology is used as a means by which globalization is effected. Works on the justification of intellectual property rights typically fall under two extremes: either they reject IPRs outright or they accept IPRs as necessary for global commerce and useful innovation. The former argue, on the one hand, that IPRs are hegemonic tools by which the developed countries in the West keep the emerging developing ones at bay or exploit the natural resources of the developing countries through what is known as biopiracy or bioprospecting. On the other hand, those who embrace IPRs usually base their arguments on the role that IPRs are necessary as a means of protecting those who have invested in creating useful innovations. Problems arise when the products protected by IPRs are carried across national borders and thus become global. In order to ensure protection afforded by IPRs across countries, a worldwide system has been created by which IPRs are protected which in many cases override the sovereignty of states. Thus it is clear that IPRs are clearly tools of globalization; one sees globalization concretely at work through the creation and enforcement of trade-related intellectual property rights across countries in the world today.

The polarized debates around IPRs have created countless cases of conflicts between those who fight for globalization and those who are against it. Chief in these debates is the ethical issue, especially when products protected by IPRs have strong impact on the livelihood and even the survival of those who depend on them. New pharmaceutical products, for example, are almost always patented, which enables the manufacturer to be able to charge very high price to cover their investments and also to earn themselves profits for their shareholders. However, when people in the poorer developing world are in need of these drugs, it is clear that there are moral issues involved. Are the pharmaceutical companies morally obligated to provide the fruits of their intellectual investments at lower cost so that they are affordable by the poor? It would strongly seem so. However, there are also cases where IPRs are justified by arguments that they are necessary as an incentive for innovation. Without effective IP protection, the life saving drugs in question might not have arisen in the first place. Furthermore, there are also cases where IPRs are used as tools for protecting the creation of those within the developing world themselves. Without workable IPR regime, it is not

quite conceivable how innovation that takes place within the developing world can even get off the ground. In fact ineffective enforcement of IPRs in the developing world has been cited as one reason for these countries remaining stagnant economically.

The present paper aims to break this impasse. The underlying issue behind the debate on patented pharmaceuticals and other products such as software or other forms of innovation is the use of IPRs as a tool for protecting intellectual creation. The intellectual content that becomes property through patents is constituted by information. Thus the issue becomes in effect how information itself is owned and how it has become a commodity. Hence it is clear that the issue depends the value one puts on the information in question. It is just not that case that information can have more or less values on its own – if the information answers to the people’s needs and desires, then naturally it is more valuable. This implies that the value a piece of information has is dependent upon context, which is mostly made up of people. Thus IPRs function when information itself has economic values and can be bought and sold. This shows that in themselves IPRs are neither positive or negative, no more than a piece of cloth sold in the market is either positive or negative. IPRs then can be used either positively or negatively. For example, when they are used to monopolize life saving drugs so that poorer people cannot afford them, then they are negative, but they can also perhaps become more positive when they are used to advance the interests of poorer people by ensuring, for example, that the plant species belonging to their natural habitats are protected, or their own intellectual creation is recognized and given due protection.

As mentioned previously, information technology plays a significant role in all this. First of all, products of information technology itself are usually protected by IPRs. Software is usually protected by copyrights. It is well known that the open source movement in software strikes a middle ground between copyright protection and commercialization on the one hand, and releasing everything onto the public domain on the other. This can be a way out of the impasse, but it needs more thorough theoretical justification, which is also an aim of this paper. Another, no less important, point is that, as the technology spreads the information around, and as information does not have values on its own as previously discussed, information technology itself stands to be used either positively or negatively too. This seems to be a come back to the old position of technological neutralism (the idea that technology is not good or bad in itself). But it is not. When one allows for all the constraints and implications associated with a technology (i.e., when a technology constrains us to behave one way or another due to the nature of that particular technology itself), there is still room for using that technology within these constraints either positively or negatively. Hence, a way is open before us and it is up to us to decide which way to go. We only need to be able to foresee, to the extent that we can, what kind of consequences there will be as a result of our choosing.

### **Acknowledgements**

Research for this paper has been partially supported by a grant from the National Research University Project, grant number HS1025A and AS569A.

# **Track IX: Surveillance, sousveillance**

## TOWARDS A HERMENEUTIC PHENOMENOLOGY OF CYBER-SPACE: POWER VS. CONTROL

ANDREAS BEINSTEINER

*Ph.D. Student*

*Institute of Philosophy*

*Leopold-Franzens-Universität Innsbruck*

**Abstract.** Since the 1990ies, regulation by program code has become an issue in theoretical reflection on computers. Michel Foucault's concepts, and, in particular, Gilles Deleuze's claim that control societies substitute disciplinary societies in the age of computers, have been popular points of reference. The present paper suggests interpreting control as a form of regulation that is essentially connected to computers: From Foucault's considerations a distinction is derived between power and control. Control is conceived as a more radical mode of regulation: a determination of possibilities of action that – as is shown by relating Foucault to Martin Heidegger – is first made possible by computer technology.

### 1. The power of code

In an article called “Soft Cities”, William J. Mitchell (2005) explores similarities and differences between traditional “real-world” space and the new, computer-generated spaces. He observes that the coded conditionals in cyberspace provide a fundamentally new mode of regulation: you cannot argue with computer programs, you cannot plead or bribe them. Lawrence Lessig (2006) refines his claim “code is law” by stating that this new form of regulation rather works through “a kind of physics. A locked door is not a command ‘do not enter’ backed up with the threat of punishment by the state. A locked door is a physical constraint on the liberty of someone to enter some space.” (p. 82) Code is a regulator in cyberspace because it defines the terms upon which a certain cyberspace environment is offered: It decides what can be said and done in that environment.

Lessig refers to Michel Foucault (1995) who had addressed the kind of regulations that become relevant in a new way in cyberspace: “Discipline and Punish” introduced the perspective that tiny corrections of space regulate by enforcing a discipline. In fact, Foucault's reflections on disciplinary power are embedded in his larger project of exploring the historical transformations that substitute sovereign power by what he calls biopower: a new kind of power that does not employ law but technology and that does not prohibit behavior but produce it. (Foucault 1998)

According to Gilles Deleuze (1995), disciplinary societies have been replaced by control societies in the age of computer technology. Alexander Galloway (2004, 2010) has characterized protocol and program code as the essential means of regulation in control societies.

## **2. Power and freedom**

According to Foucault, to exercise power means to structure the possible field of action of others. By doing so, these individuals are transformed into subjects, where the word subject has two meanings: to be subject to someone else's domination, and to be tied to one's own identity.

Foucault (2002) emphasizes that power can only be exercised over free subjects. A subject is free insofar it is not absolutely self-identical or determined. In the extreme case where power constraints action absolutely or physically, both power and freedom disappear: "slavery is not a power relationship, when man is in chains." (p.221) I suggest conceiving control as such a form of regulation that goes beyond power and erases freedom.

While the absence of physical determination seems to be a necessary condition for freedom, it is not a sufficient one. Since it does not seem adequate to suppose a kind of metaphysical autonomy in Foucault's conception of the individual, we turn to the relations that Hubert Dreyfus (2003) has established between the concepts of Foucault and Martin Heidegger for a deeper understanding of how to conceive the sources of freedom. According to Dreyfus, Heidegger's question – how things have turned into objects in modernity – is complemented by Foucault's question – how individuals have been turned into subjects. This allows connecting Heidegger's concept of Being with Foucault's concept of power. Since one's goals and horizons of meaning arise from one's background understanding that Heidegger calls the clearing of Being, exercising power over a certain individual (to influence his/her possibilities of action) is possible by shaping this clearing. A subject is constituted by the corresponding understanding of Being, and the more static this understanding is, the closer to absolute self-identity is the subject. Thus freedom can be grasped as hermeneutic oscillation – as a condition where various understandings are suspending and balancing each other.

## **3. Materiality as a source of freedom**

According to Heidegger, the understanding of Being has always been influenced by technological artefacts and vice-versa. A tool suggests what it is to be used for: Heidegger's (1995) prominent example is the hammer, which is embedded in a structure of "in-order-to"-relations and refers to goals, practices and other tools.

In contrast to tools, whose materiality disappears into their usability, works of art emphasize their materiality. By doing so, they expose a fundamental gap between the material sphere and the conceptual sphere. Heidegger (2008) conceives this as a struggle between earth and world. The artwork's materiality cannot be exhaustively interpreted with one conceptual frame, thus it steadily keeps evoking new interpretations. This is how materiality provides a source of freedom. Also tools, due to their materiality, may

be abused or used in different ways that were not intended originally. Addressing what he calls the “designer fallacy”, Don Ihde (2009) has examined such non-intended usages of technologies. Ihde’s argument against the possibility to design in advance a tool’s usage relies on the tool’s materiality.

#### 4. Cyberspace as the congruence of material and conceptual

For a long time theology and science employed god’s order of creation or the capacity of human reason to bridge the gap between the conceptual and the material sphere. (Heidegger 2008) The task of metaphysics was to provide narratives that justified the adequacy of a certain vocabulary for describing reality. Nietzsche’s “death of god” is nothing but the acknowledgement that there is not one single conceptual system that adequately describes reality. The “post-modern” call for conceptual pluralism is a consequence from this insight.

In cyberspace environments, however, the productive tension between the material and the conceptual is erased: The programmer is the god who creates this reality, and the respective program code is really an adequate description of this reality. Conceptual and material sphere coincide in cyberspace. A gun in a 3D shooter game is nothing but a gun and a buy-with-one-click-button in an online shop is nothing but a buy-with-one-click-button. The “designer fallacy” argument does not hold in cyberspace. And thus, as agents in a cyberspace environment, we are 100% self-identical subjects. According to my suggestion, this is what control is about.

#### References

- Deleuze, Gilles (1995): *Negotiations, 1972-1990*. New York: Columbia University Press.
- Dreyfus, Hubert (2003): 'Being and Power' Revisited. In Milchman, Alan & Rosenberg, Alan: *Foucault and Heidegger: critical encounters* (pp. 31-54). Minneapolis: University of Minnesota Press.
- Foucault, Michel (1995): *Discipline and punish: the birth of the prison*. New York: Vintage.
- Foucault, Michel (2002): The Subject and Power. In Dreyfus, Hubert and Rabinow, Paul: *Michel Foucault: Beyond Structuralism and Hermeneutics* (pp. 208-226). New York: Harvester Whitesheaf.
- Foucault, Michel (1998): *The Will to Knowledge*. London: Penguin Books.
- Galloway, Alexander R. (2004): *Protocol. How Control exists after Decentralization*. Cambridge, Massachusetts: MIT Press.
- Galloway, Alexander R. (2010): Networks. In Mitchell, W.J.T. and Hansen, Mark (Eds.): *Critical terms for media studies* (pp. 281-296). Chicago: University of Chicago Press.
- Heidegger, Martin (2008): *Basic Writings*. New York: Harper Collins.
- Heidegger, Martin (1995): *Being and Time*. Oxford: Blackwell.
- Ihde, Don (2009): The Designer Fallacy and Technological Imagination.”In Vermaas, Pieter E. et al. (Eds.): *Philosophy and Design. From Engineering to Architecture* (pp. 51-59). Springer
- Lessig, Lawrence (2006): *Code version 2.0*. New York: Basic Books.
- Mitchell, William J. (2005): *City of Bits*. Cambridge, Massachusetts: MIT Press.

## THE WIKILEAKS LOGIC

JEAN-GABRIEL GANASCIA  
*LIP6 – University Pierre et Marie Curie*  
*4, place Jussieu, 75005, Paris, France*  
*Jean-Gabriel.Ganascia@lip6.fr*

**Abstract.** WikiLeaks has focused the attention of the media during a few weeks by the end of 2010. The diplomacy of the United-State of America has been called into question. Modern democracies are hampered; as sovereign states, they are now facing a novel dilemma. This paper constitutes an attempt to understand this evolution by seriously considering the WikiLeaks project not as a simple media strategy, but as the possible kickoff of a totally new way doing politics, in a perfect transparency, without secrecy nor hidden issues. Our purpose here is both to show how information technologies, of which WikiLeaks is a sub-product, contribute to transform the traditional political forms and how the notion of “sousveillance” helps us to apprehend these evolutions.

### 1. A Few Recent Facts

WikiLeaks has focused the attention of the media during a few weeks by the end of 2010 and, previously, during the summer and the autumn. The diplomacy of the United-State of America and of some other countries has been called into question by what people called the *Cablegate*, by analogy to the *Watergate*. Let us remember that 250,000 of secret telegrams containing embarrassing information about American, European and Middle-East foreign policies were divulged to newspapers by the WikiLeaks organization. Modern democracies, and especially the United-States of America, were hampered. The main argument they developed against WikiLeaks was formal: it concerned the danger that was posed to those whose name had been explicitly mentioned in the cables. However, it clearly appeared that, for those sovereign states, the question is not only just saving life of a few people: they are now facing a novel dilemma. On the one hand, last few years many democracies opened public data to all citizens (Obama 2009). On the other hand, states are always used to deal with many matters, especially in the diplomatic area, either in secrecy, or, at least, in a discrete way. As a consequence, they can't easily accept the divulgation of top secret informations. In brief, the aspiration to a total transparency, that many of our contemporaries share, modifies the rules of government, while WikiLeaks shows the limits of officially proclaimed public transparency.

## **2. A New Ideal of Transparency**

With the recent developments of information technologies a new ideal of total transparency seems to be born. Note that, by itself, the ideal of total transparency is not new. It already existed in the 19<sup>th</sup> century (Benjamin 1934). The use of glasses in the architecture, for instance the “Crystal Palace” that was built for the London Universal Exhibition in 1851, reflected this ideal.

A few years before, in the end of the 18<sup>th</sup> century, Jeremy Bentham had described an architecture for surveillance designed to ensure a total transparency (Bentham 1838). Called the Panopticon, it was a model for prisons, factories, hospitals, etc., that have been conceived to make individuals totally visible to their guards, while these ones were invisible to them. The goal of transparency was again to facilitate education, surveillance, care, etc., which enhanced the role and the situation of authority holders. By contrast, the new transparency that is encouraged today is individual and not institutional. It is directed towards and against the authority holders, which are permanently under the cameras. For instance, the policemen are continuously filmed. The professors, physicians, lawyers, politicians etc. are permanently evaluated, etc. The concept of “sousveillance” that was introduced by Steve Mann well characterizes this new form of transparency (Mann 2003). This neologism forged by analogy and opposition to the word surveillance, means that the watcher is situated below (“sous” in French) the authority, while in case of surveillance he is situated above.

## **3. The Horizon of WikiLeaks**

To understand the horizon of WikiLeaks, let us first note that Julian Assange, the promoter and editor in chief of WikiLeaks, was initially a computer scientist who first worked on cryptography. So doing, he adopted an atypical posture. While almost all the cryptographers work for armies, secret services or banks, he developed cryptographic tools for people. His idea was to make everybody able to hide information to the authorities (state, company, etc.).

Now, with WikiLeaks, Julian Assange proposes to render publicly available all information about authorities. He proposes creating “open governments” where all data about the government and the public decisions would be worldwide accessible to everybody. The underlying idea of a perfect collective transparency seems to justify his action, which somehow refutes his first attitude of privacy protection.

## **4. Limits of the Generalized Sousveillance**

The utopia of a generalized sousveillance, i.e. of a sousveillance extended to the overall society, that excludes surveillance, faces an inherent contradiction: the authorities are made of individuals, who, as such, need to be protected, which becomes impossible because of the exclusion of surveillance.

Without going deeply in the exploration of this first contradiction, consider now the extension of the sousveillance regime to the overall worldwide society. It faces at least two types of limitations, some being intrinsic, others extrinsic.

The main intrinsic limitation is due to our cognitive abilities that are too limited to permit to observe and to assimilate all the information we have at our disposal. As a consequence, we spontaneously filter the information flows and we focus our attention on the most prominent facts. But, we do not decide by ourselves what criteria are adopted to qualify the prominence. Most of the time, this is decided by people who manipulate us by distracting our attention.

The second type of limitation is extrinsic in the sense that it is not an own limit of the regime of sousveillance itself, but it is due to foreign factors. Specifically, nothing prohibits the coexistence of a generalized regime of sousveillance with multiple regimes of surveillance. For instance, NGOs or big multinational companies may continue to gather and exploit data; they even can take advantage of free public data to extract useful knowledge for the sake of their own interest, without any respect of privacy.

## 5. The Failure of the Wikileaks Ideal

Despite the attacks to which it was submitted and the fact the Julian Assange has been jailed, WikiLeaks is undoubtedly very popular nowadays. There even exist attempts to build more or less specialized clones of WikiLeaks in many places all over the world. However, the original Assange project seems to have failed. The causes of this failure are directly related to the limitations of the generalized sousveillance regime that were expressed in the previous paragraph.

First of all, Julian Assange wanted to freely disseminate data allowing every citizen to get any information he wanted, when he wanted. However, during the Cablegate, WikiLeaks didn't freely divulge the 250,000 diplomatic telegrams he had; he sent them to well established newspapers that had to filter, anonymize the messages and dramatize their publication, with appropriate comments and advertisements.

Another failure of the WikiLeaks project is due to the project itself, which was supposed to free people from any kind of authorities. However, it clearly appears that WikiLeaks has now become a new authority, which plays a role symmetrical to other more traditional authorities, as states or NGOs and companies. Julian Assange himself acts in his own organization without any real transparency, which shows the limitation of the generalized sousveillance principle as it was promoted by WikiLeaks.

## References

- Benjamin, W. (1934), *Selected Writings*, Volume 2, 1927-1934 Translated by Rodney Livingstone and Others Edited by Michael W. Jennings, Howard Eiland, and Gary Smith
- Bentham, J. (1838), Panopticon or the Inspection House, *The Work of Jeremy Bentham*, volume IV, 37-172
- Mann, S., Nolan, J., Wellman, B. (2003), Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments, *Surveillance & Society* 1(3): 331-355, <http://www.surveillance-and-society.org> - <http://wearcam.org/sousveillance.pdf>
- Obama, B., (2009), Transparency and Open Government, *Memorandum for the Heads of Executive Departments and Agencies*, The White House, Washington, USA, [http://www.whitehouse.gov/the\\_press\\_office/Transparency\\_and\\_Open\\_Government/](http://www.whitehouse.gov/the_press_office/Transparency_and_Open_Government/)

## DEMOCRACY 2.0 - HOW THE WEB MAKES REVOLUTION

ANIS NAJAR

*LIP6, Pierre and Marie Curie University*

*4, Place Jussieu, 75005, Paris, France anis.najar@lip6.fr*

**Abstract.** “Whoever controls the information owns the power”. Many scientists and philosophers have been interested in analyzing the relationship between information and power within the society and they all argued that a kind of dependency exists between the control of information and the political power. In this paper, we propose to analyze this dependency from a structuralistic point of view by assuming that changes in the information schema of the society would necessarily produce changes in the power schema, characterizing by this way the concepts of surveillance and sousveillance. We suggest examining these changes on two levels, the structure of the information schema and the nature of information, by taking as a study case the Tunisian popular revolution in which information technology have played a significant role.

### 1. Introduction

From a structuralistic point of view, we can model the information society as entities exchanging information in some pattern that we will refer to as information schema. Similarly, we will call power schema, the one representing the balance of power between the entities within the society. By neglecting other socioeconomic factors, we can say that the power schema is somehow characterized by the information schema. Therefore, it is reasonable that a revolution in the latter produces a revolution in the former. To illustrate these aspects, we take as a study case the Tunisian popular revolution that we consider as a logical consequence of the anterior revolution of Information Society. Indeed, yet five years ago, the World Summit on the Information Society held in Tunisia reflected the contradiction in the dictator's policy towards Information Technology. At the same time, he was promoting its use and censoring its access. In effect, he was not suspecting at that time, that five years later he would be overthrown by what he was the most proud of, i.e. Information Technology. In the following, we try to analyze this revolution on two levels, namely the structure of the information schema and the nature of information itself.

## **2. Informational Revolution**

### **2.1. STRUCTURAL LEVEL**

Based on the concept of Panopticon introduced by Jeremy Bentham in 1785 (Bentham 1838), Michel Foucault (Foucault 1975) described the classical schema of surveillance in a society as a hierarchical organization, in which the state controls the information either in its dissemination through the media and education or its collection through intelligence. This schema also defines the classical power schema as a vertical organization, the state at the top and the people at the bottom. Besides, censorship has often been the classical way of controlling the information in such configuration. Since several years ago, Internet has substantially transformed the information schema which progressively took the form of the World Wide Web structure, that of network. This reversed the power schema in a way that balanced the power relationship between the state and the people by promoting transparency of information and democratization of power. This schema coincides with the architecture of Catopticon introduced by Jean-Gabriel Ganascia (Ganascia 2009) in order to describe the structure of “sousveillance”, in opposition to Bentham's Panopticon. Sousveillance has been defined by Steve Mann (Mann 2003) as the acquisition by people of information technology so they can use it against their keepers.

During Tunisian revolution, we observed a real showdown between the people and the government, especially through social networks that have been a real staging ground for the demonstrations. The advantage provided by the internet can be explained by several reasons. First, notions such as community and sharing that have been developed through social networks like Wikipedia, Facebook and Twitter have created a kind of proximity between people and strengthened their solidarity. Second, the distribution aspect of networks and speed of information propagation (small world effect) make social networks a very effective offensive tool. For example, the worldwide cyber-activist organization known as Anonymous launched an operation called #OpTunisia against the Tunisian Internet Agency servers paralyzing several government web sites. Moreover, the great demonstration that led to the departure of the dictator has been organized via Facebook overnight just after his last speech. Third, this structure is robust against targeted attacks because of the absence of “leaders”. Finally, it is effective against censorship because it is always possible to introduce information from a part of the network.

### **2.2. SEMANTIC LEVEL**

The second aspect of change in the information society has been made in the nature of information contents. For some time indeed, the multimedia, especially video is being increasingly important within the information exchanged over the Internet. We could explain this by several reasons. First, the constraints of formalization and formulation downsized the previously privileged position of texts, leaving the ground for videos which appeared to be a more effective mode of information circulation in terms of quickness and straightforwardness. Second, in addition to the fact that image is semantically richer than text; it is also much closer to the human's mental representation;

so it allows a better effect on the mental image, which gives it more impact in information transmission.

All these factors contributed to the success of video particularly through video-blogging and gave birth to a new kind of media, which is the collaborative journalism, where everyone contributes to the spreading of information. Furthermore, many news TV channels, when they were not allowed to directly cover events, had no other choice than collecting and sorting amateur videos provided by protestors in order to broadcast them afterwards.

### 3. Counter-Revolution

Even though the network structure, as we exposed, is resistant against attacks, there is still one kind of attack that is effective against information networks and which takes advantage of its foregoing characteristics, that is propaganda. That was an essential tactical point that let the former regime to launch a counter-revolution by changing its behavior in a second time from censorship to disinformation. It seems that they understood that they would be more able to control information by fabricating it rather than by blocking it. For example, in just a brief delay after the censorship has been lifted on the internet, multiple Facebook pages have been created to turn the opposition parties against each other and the Ministry of Interior created an official page to make propaganda. In a few hours, Facebook has been flooded by a huge quantity of rumors about criminals and snipers shooting people outside so that terror led people to not think rationally and they didn't trust any information anymore. By this way, the government created chaos and paralyzed the network.

In the same way, image has also been used in the counter-revolution. For the same reasons cited above it has been a very effective tool of manipulation. For example, in attempt to discredit protestors, the government staged several acts of violence and spread them on the internet so that a lot of people called to stop demonstrations.

### References

- Bentham, J. (1838), *Panopticon or the Inspection House*, *The Work of Jeremy Bentham*, volume IV, 37-17
- Foucault, M. (1975), *Surveiller et punir*, Gallimard, Paris, France, p. 252 – In English *Discipline and Punish*, trans. A. Sheridan. (1977) New York: Vintage
- Ganascia J.-G. (2009), "The Great Catopticon", in *Proceedings of the 8th International Conference of Computer Ethics Philosophical Enquiry (CEPE)*, 26-28 June 2009, Corfu, Greece
- Mann, S., Nolan, J., Wellman, B. (2003), *Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments*, *Surveillance & Society* 1(3): 331-355, <http://www.surveillanceand-society.org> - <http://wearcam.org/sousveillance.pdf>

## NEGATIVE SOUSVEILLANCE

CARSON REYNOLDS

*University of Tokyo, Department of Creative Informatics*  
*carson@k2.t.u-tokyo.ac.jp*

**Abstract.** Recent catastrophes have increased the desire to get rapid information about infrastructure such as power and services and not necessarily from the people providing these services. While news sources seek to provide such information, they are biased toward providing information that increases reader or viewer interest. Sousveillance is appropriate in these cases and here we describe an unusual method for such observation, which we call negative sousveillance. This is observing which systems or services disappear in a time of catastrophe and reporting on their disappearance.

### 1. What Disappeared?

Mann's development of "watchful vigilance from underneath" is useful in cases in which the surveilled feel that information may be used to harm them. But what of the special case in which the disenfranchised feel that information is being withheld from them?

Amid the recent earthquake, tsunami, and nuclear power crises of Japan in 2011, several individuals have expressed to me the feeling that they "are not being told everything." Indeed, Wikileaks's (Pilger, 2010) recent diplomatic cable archive documents the extent that governments and organizations routinely keep politically delicate details out of the public eye.

Negative databases (Esponda, 2006), on the other hand, are designed to solve a different problem altogether. That is the keeping records which if stolen do not reveal the identity of individuals. Negative databases achieve this by storing the complement of the set of what is being tracked. Essentially the database shows what isn't of concern.

The work of Trevor Paglen, involves long-distance photography and data analysis to document secret installations. Extending his approach the negative intelligence gatherer would seek to understand what websites, infrastructure systems, environmental sensors or documents have become unavailable.

The negative sousveillance concept then is to record, track, or infer what isn't there. This essentially suggests a two-stage process. The first step is citizens or activists to survey or map infrastructure systems or environmental status. Paulos, Honicky, and Hooker (2009) showed how urban populations could use mobile phones as dense environmental sensors for citizen science. Analogously, Bonanni et al. (2010) have created a system for tracking and account supply chains and their environmental effects.

Project such as OpenStreetMap have already sought to create public domain maps of the physical world. The second step is to record what has disappeared.

The approach is broadly applicable. Those interested in digital image manipulation can keep a delta showing how an image is gradually altered over time through the addition of watermarks or removal of figures from the scene. Those interested in network systems can track network outages due to disasters or *kill switches*, which would be used by governments to limit internet access (Cowie, 2011).

The practices of negative information gatherers in some cases would be similar to those of network security professionals. They might proceed by using tools such as *nmap* to scan various network services and store them into a database (Lyon, 2009). As services disappear they would then be listed in the far more interesting negative database. Those interested in environmental sensors may either try to gain access to the sensor data or distribute their own environmental sensor network. When nodes in such a network stop responding further investigation is warranted. It may be that the network node needs to be replaced, that it has been tampered with, or destroyed by environmental causes. But the absence of information is just as interesting as steady broadcast.

The anticipatory step of documenting infrastructure before it disappears is also useful in disaster situations when officials may be inundated with requests for information. I believe the question “is X inoperative” is an easier question to answer to than “what type of X exist and are they inoperative?” With careful foresight the negative database may be able to answer both questions without relying officials or outside organizations for details.

## 2. Skepticism & DIY Authority

The feeling of powerless that comes from lack of information can be alleviated by the realization that you yourself can gather information. While news sources, corporate press releases, and government agencies often have access to expert assessment I think it is fair to question whether such experts have biases. For instance, news outlets may err on the side of sensationalism to stir up concern about a recent event; corporations may time announcements to minimize the impact of bad news (Gross, 2004), or agencies may try to minimize widespread panic at the expense of accurate information.

One interesting aspect of *DIY* infrastructure, environment, or network monitoring is that those affected can collect and analyze details that affect them. When objects disappear from view instead of entering a memory hole they are instead specially noted as they are entered into a negative database. It is our hope that less will escape the notice of those willing to do the legwork involved in becoming authorities themselves.

## References

- Bonanni, L., Hockenberry, M., Zwarg, D., Csikszentmihalyi, C., & Ishii, H. (2010). Small business applications of sourcemap. Proceedings of the 28th international conference on Human factors in computing systems - CHI '10 (p. 937). New York, New York, USA: ACM Press. doi: 10.1145/1753326.1753465.
- Cowie, J. (2011). Egypt Leaves the Internet. Retrieved from <http://www.renesitys.com/blog/2011/01/egypt-leaves-the-internet.shtml>.

- Esponda, F., Ackley, E., Helman, P., Jia, H., & Forrest, S. (2006). Information Security. (S. K. Katsikas, J. Lpez, M. Backes, S. Gritzalis, & B. Preneel, Eds.) Information Security, Lecture Notes in Computer Science, 4176, 72-84. Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/11836810.
- Gross, D. (2004). Friday Night Blights. Slate. Retrieved from <http://www.slate.com/id/2106864/>
- Lyon, G. F. (2009). Nmap Network Scanning: The Official Nmap Project Guide to Network Discovery and Security Scanning. Retrieved March 16, 2011, from <http://portal.acm.org/citation.cfm?id=1538595>
- Mann, S., (1998), 'Reflectionism' and 'diffusionism': new tactics for deconstructing the video surveillance superhighway, *Leonardo*, **31**(2): 93–102.
- OpenStreetMap Foundation. (2011). OpenStreetMap. Retrieved from <http://www.openstreetmap.org/>
- Paglen, T. (2011). Visual Projects. Retrieved on March 14<sup>th</sup>, 2011 from <http://www.paglen.com/pages/projects.htm>
- Pilger, J. (2010). Why WikiLeaks must be protected. *New Statesman*, 139(5015), 18.
- Paulos, E., Honicky, R., & Hooker, B. (2009). No Title. Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City (pp. 414-436). doi: 10.4018/978-1-60566-152-0.ch028.

## GOVERNMENT APPROACHES FOR MANAGING ELECTRONIC IDENTITIES OF CITIZENS – EVOKING A CONTROL DILEMMA?

STEFAN STRAUSS

*Austrian Academy of Sciences, Institute of Technology Assessment (ITA)*

*Strohgasse 45/5, A-1030, Vienna, Austria*

*sstrauss@oeaw.ac.at*

**Abstract.** Governments world-wide introduce electronic identity systems to adapt the process of citizen identification to the needs of the information society. These innovation processes primary aim at improving e-government services, but imply further societal and political objectives. The emergence of identity management represents a demand for (re)gaining control over personal data in virtual environments. Compared to predominating security goals, privacy aspects are often neglected and not sufficiently implemented. The analysis from a privacy perspective shows that the current situation of governmental e-ID can be described as a control dilemma: despite of its aim to (re)gain control, the e-ID could ironically even foster a further loss of control over individual privacy. As a consequence, an e-ID system itself might turn into a sort of amplified surveillance interface. In this regard, the e-ID could become a synonym for a panoptic instrument of power. The e-ID example refers to the major challenge of enhancing governmental transparency for individuals and the public sphere to compensate a further growth of information asymmetries and imbalanced control over personal information between citizens and governments.

Information and communication technologies continually pervade everyday life and change the dynamics of data processing and information handling in many respects. Significant increases in personalized services and social interactions over web 2.0 applications inevitably entail further growth of digital data, aggravating individuals in controlling personal information and protecting their privacy. The convergence of analog and digital environments further accelerates these trends. The increasing relevance of electronic identity management (IDM) as an important field of research in the information society (Halperin/Backhouse 2008) is a prominent example for this convergence. While many different IDM concepts exist, especially national governments made remarkable efforts in recent years to introduce electronic ID cards for supporting online public services; primary objectives are improving security and unifying identification and authentication procedures in e-government.

Identification is a core function of governments and thus the creation of national e-ID systems implies far-reaching societal transformations (Aichholzer/Strauß 2010) that contribute “to alter the nature of citizenship itself” (Lyon 2009). Hence, e-ID is more than an identification device; it becomes a policy instrument, and the focus more and

more shifts from being a “detecting” tool to an “effecting” tool; i.e., an instrument not only to support administrative procedures such as ascertaining identity in public services but to enable services and to impact societal and political objectives (Bennett/Lyon 2008). Inter alia in EU information society policies the vision is to set up a “pan-European infrastructure for IDM in support of a wide range of e-government services” (CEN 2004); and introducing e-IDMS also aims at fighting identity fraud and terrorism (CEN 2004). Privacy is obviously of vast importance but plays a rather implicit role while security issues predominate. Although e-ID introduction is not to be seen as a consequence of the 9/11 tragedy, this strong security focus was catalyzed in some respect by it (Bennett/Lyon 2008). E-ID cards “have become the tool of choice for new forms of risk calculation” and enable a “mode of pre-emptive identification” (Lyon 2009). History offers many examples for social discrimination and population control, drastically illustrating the strong relations between identification and surveillance (Bennett/Lyon 2008; Lyon 2009). But IDM is not inherently a privacy threat. Whether an e-IDMS becomes an instrument of surveillance or not naturally depends on the concrete system implementation and its surrounding framework. Properly designed with respect to privacy enhancement, e-IDMS might contribute to informational self-determination; i.e., proactively support individuals in handling their different identities in different contexts and controlling their personal data (Clauß et al 2005), which is the very idea of IDM.

However, current e-ID card schemes only rudimentarily include privacy mechanisms and do not correspond to privacy-enhancing IDM (Naumann/Hobgen 2009). Particular problems are insufficient implementations of anonymity and pseudonymity, undermining the concept of unlinkability, which is essential to prevent “privacy-destroying linkage and aggregation of identity information across data contexts” (Rundle et al 2008). The growing amount of personal data due to further trends towards pervasive computing environments intensifies these problems as identity never shrinks (Pfitzmann/Borcea-Pfitzmann 2010). The increasing visibility of identification mechanisms entails a sort of shadow<sup>22</sup>. This “identity shadow” facilitates data linkage and de-anonymization (Strauß 2011). Surveillance tendencies and predominant security objectives in the e-ID development imply further frictions. Combined with the evident danger of function creep, i.e., a purpose extension of e-ID usage, this could lead to the advent of a ubiquitous IDM infrastructure entailing further privacy threats. The current situation can be described as a control dilemma: while the increasing role of IDM represents “a demand to regain control over personal data flowing in digital environments”, the creation of governmental e-IDMS to fulfill this demand could ironically even foster a further loss of control over individual privacy (Strauß 2011).

In this sense, an e-IDMS has several similarities to Foucault’s (1977) interpretation of the panopticon “as a generalizable model of functioning; a way of defining power relations in terms of the everyday life of men”. Social control becomes automated as the algorithms of the system define the way one’s identity is treated, i.e., the degree of service provision based on automated categorization. The trap of visibility (Foucault 1977) here is the increasing ID-obligation triggered by the e-IDMS. While the system becomes more and more visible, its functioning becomes further blurred for individuals. They have to reveal their ID without knowledge about whether and for what purpose it is used - analog to the uncertain presence of the guard in the watchtower. Consequences

---

<sup>22</sup> In recognition of Alan Westin: Privacy and Freedom, 1967 and the term “Data Shadow”.

would be self-censorship and limited individual freedom because “without transparency, one cannot anticipate or take adequate action“ (Hildebrandt 2008).

The control dilemma highlights the demand for more effective privacy concepts and control mechanisms, enabling citizens and the public sphere in controlling proper and legal data usage. One crux is the system inherent realization of anonymity and pseudonymity; and, related, a thorough data minimization, e.g., addressed by already arising approaches (e.g., <http://vanish.cs.washington.edu>) for an expiration date of digital data (Mayer-Schönberger 2009). However, their practicability is limited and they cannot solve the problem of information asymmetries between the governed and those who govern. Thus, the major challenge is to compensate this imbalanced control over personal information by enhancing governmental transparency for individuals and the public sphere.

## References

- Aichholzer, G. & Strauß, S. (2010). Electronic Identity Management in e-Government 2.0: Exploring a System Innovation exemplified by Austria. *Information Polity* 15(1-2), 139-152.
- Bennett, C. J. & Lyon, D.(2008). *Playing the identity card - surveillance, security and identification in global perspective*. London and New York: Routledge.
- Clauß, S., Pfitzmann, A., Hansen, M., Herreweghen, E. V. (2005). Privacy-Enhancing Identity Management, No. issue 67, Institute for Prospective Technological Studies (IPTS).
- Comité Européen Normalisation - CEN (2004). CEN/ISSS Workshop eAuthentication - Towards an electronic ID for the European Citizen, a strategic vision, Brussels.
- Foucault, M. (1977). *Discipline and punish: the birth of the prison*, trans. A Sheridan, London: Penguin.
- Halperin, R. & Backhouse, J. (2008). A roadmap for research on identity in the information society. *Identity in the information society* 1(1), 71-87.
- Hildebrandt, M. (2008). Profiling and the rule of the law. *Identity in the information society* 1(1), 55-70.
- Lyon, D. (2009). Identifying citizens - ID cards as Surveillance. Cambridge: Polity Press.
- Mayer-Schönberger, V. (2009). Delete: The Virtue of Forgetting in the Digital Age. Princeton: University Press.
- Naumann, I., Hobgen, G. (2009). Privacy Features of European eID Card Specifications: European Network and Information Security Agency – ENISA.
- Pfitzmann, A. & Borcea-Pfitzmann, K. (2010). Lifelong Privacy: Privacy and Identity Management for Life. In: Bezzi, M. et al (Eds): *Privacy and Identity Management for Life*, Proc. of the 5<sup>th</sup> Int. PrimeLife/IFIP Summer School, IFIP AICT Vol. 320, (pp.1-17). Heidelberg: Springer LNCS.
- Rundle, M., Blakley, B., Broberg, J., Nadalin, A., Olds, D., Ruddy, M., Guimaraes, M. T. M., Trevithick, P. (2008). At a crossroads: "Personhood" and digital identity in the information society, No. JT03241547, OECD.
- Strauß, S. (2011). The Limits of Control – (Governmental) Identity Management from a Privacy Perspective. In: Fischer-Hübner, S., et al (Eds), *Privacy and Identity Management for Life*, Proc. of the 6<sup>th</sup> Int. PrimeLife/IFIP Summer School – revised selected papers, IFIP AICT Vol. 352, (pp.206-218). Heidelberg: Springer LNCS.

**Track X:  
SIG Track – Machines and  
Mentality**

## MORAL EMOTIONS FOR ROBOTS

RONALD C. ARKIN

*Mobile Robot Laboratory, Georgia Institute of Technology  
85 5<sup>th</sup> ST NW, Atlanta, GA 30332 U.S.A.*

As robotics moves toward ubiquity in our society, there has been only passing concern for the consequences of this proliferation (Sharkey, 2008). Robotic systems are close to being pervasive, with applications involving human-robot relationships already in place or soon to occur, involving warfare, childcare, eldercare, and personal and potentially intimate relationships. Without sounding alarmist, it is important to understand the nature and consequences of this new technology on human-robot relationships. To ensure societal expectations are met, this requires an interdisciplinary scientific endeavor to model and incorporate ethical behavior into these intelligent artifacts from the onset, not as a post hoc activity. We must not lose sight of the fundamental rights human beings possess as we create a society that is more and more automated. One of the components of such moral behavior, we firmly believe, involves the use of moral emotions.

Haidt (2003) enumerates a set of moral emotions, divided into four major classes: Other- condemning (Contempt, Anger, Disgust); Self-conscious (Shame, Embarrassment, Guilt); Other-Suffering (Compassion); Other-Praising (Gratitude, Elevation). Allen et al (2006) assert that in order for an autonomous agent to be truly ethical, emotions may be required at some level: “While the Stoic view of ethics sees emotions as irrelevant and dangerous to making ethically correct decisions, the more recent literature on emotional intelligence suggests that emotional input is essential to rational behavior”. These emotions guide our intuitions in determining ethical judgments, although this is not universally agreed upon (Hauser, 2006). From a neuroscientific perspective, Gazzaniga (2005) states: “Abstract moral reasoning, brain imaging is showing us, uses many brain systems”, where he identifies the locus of moral emotions as being located in the brainstem and limbic system.

The relatively young machine ethics community has focused largely to date on developmental ethics, where an agent develops its own sense of right and wrong in situ. In general, these efforts largely ignore the moral emotions as a scientific basis worthy of consideration. Nonetheless, considerable research has been conducted regarding the role of emotions in robotics, including work in our laboratory over the past 20 years (Arkin, 2005; Moshkina et al 2011). Far less explored in robotics is the set of moral secondary emotions, and their role in robot behavior and human-robot interaction. One example is where De Melo et al (2009) have demonstrated that the presence of moral affect in human-robot interaction is both discernible and enhances the interplay between humans and robot-like avatars.

Our own research (Arkin and Ulam, 2009) in the moral affective space research is illustrated by the use of guilt being incorporated into an ethical robotic software architecture designed for lethal military applications. Guilt is “caused by the violation of moral rules and imperatives, particularly if those violations caused harm or suffering to others” (Haidt, 2003) and is recognized as being capable of producing proactive, constructive change (Tangney et al, 2007). The specific architectural component we have

implemented, referred to as the ethical adaptor, incorporates Smits and De Boeck's (2003) mathematical model of guilt, which is used to proactively alter the behavior of the robotic system in a manner that will lead to a reduction in the recurrence of an event which was deemed to be guilt-inducing. In our initial application, this focuses on the deployment of lethal autonomous weapons systems in the battlefield, with respect to unexpectedly high levels of battle damage. Simulation results demonstrate the ethical adaptor in operation.

For non-military applications, we hope to extend this earlier research into a broader class of moral emotions, such as compassion, empathy, sympathy, and remorse, particularly regarding the use of robots in elder or childcare, in the hopes of preserving human dignity as these relationships unfold in the future. There is an important role for artificial emotions in personal robotics as part of meaningful human-robot interaction, and having worked with Sony Corporation on their AIBO and QRIO entertainment robots (Arkin, 2005), and Samsung for their humanoid robots (Moshkina et al, 2011), it is clear that value exists for their use in establishing long-term human-robot relationships.

There are, of course, significant ethical considerations associated with this use of artificial emotions in general, and moral emotions in particular, due in part to their deliberate fostering of attachment by human beings to non-human artifacts. This is believed to promote detachment from reality by the affected user (Sparrow, 2002). While many may view this as a benign, or perhaps even beneficial effect, not unlike entertainment or video games, it can clearly have deleterious effects if left unchecked, hence the need for incorporating models of morality within the robot itself.

## Acknowledgements

This research was supported under Contract #W911NF-06-1-0252 from the U.S. Army Research Office. The author would also like to acknowledge Patrick Ulam for his contribution in software development for this project.

## References

- Allen, C., Wallach, W., and Smit, I., (2006). "Why Machine Ethics?" *IEEE Intelligent Systems*, July.
- Arkin, R.C., (2005). "Moving Up the Food Chain: Motivation and Emotion in Behavior-based Robots", in *Who Needs Emotions: The Brain Meets the Robot*, Eds. J. Fellous and M. Arbib, Oxford University Press.
- Arkin, R.C. and Ulam, P., (2009). "An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions", *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09)*, Daejeon, KR.
- De Melo, C., Zheng, L. and Gratch, J., (2009). "Expression of Moral Emotions in Cooperating Agents". *9th International Conference on Intelligent Virtual Agents*, Amsterdam.
- Gazzaniga, M., (2005). *The Ethical Brain*, Dana Press.
- Haidt, J. (2003). "The Moral Emotions", in *Handbook of Affective Sciences*, Oxford Press.
- Hauser, M., (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*, ECCO, HarperCollins, N.Y., 2006.

- Moshkina, L., Park, S., Arkin, R.C., Lee, J.K., Jung, H., (2011). "TAME: Time-Varying Affective Response for Humanoid Robots", *International Journal of Social Robotics*.
- Sharkey, N. (2008). "The Ethical Frontiers of Robotics", *Science*, (322): 1800-1801.
- Smits, D., and De Boeck, P., (2003). "A Componential IRT Model for Guilt", *Multivariate Behavioral Research*, Vol. 38, No. 2, pp. 161-188.
- Sparrow, R., (2012). "The March of the Robot Dogs", *Ethics and information Technology*, Vol. 4(2).
- Tangney, J., Stuewig, J., and Mashek, D., (2007). "Moral Emotions and Moral Behavior", *Annu. Rev. Psychol.*, Vol.58, pp. 345-372.

## ON DEEPLY UNCONSCIOUS INTENTIONAL STATES

KONSTANTINE ARKOUDAS  
*Telcordia Research*  
*Piscataway, NJ, USA*

In this note I will argue against the thesis that humans are equipped with computational structures and algorithms that are unconsciously used for logical reasoning. This thesis represents the received view in cognitive science, particularly in the psychology of reasoning. According to it, the processes by which people reason are unconscious and therefore inaccessible to introspection. The unconsciousness that these cognitive scientists allege is deep. Unconscious mental states of this form are not like the preconscious states of Freud, such as beliefs that can be ascribed to me when I am in dreamless sleep. For instance, when I am asleep I continue to believe that the second world war ended in 1945, even though I do not consciously entertain that belief during that time. The belief is preconscious; even though it is not conscious most of the time, I can easily bring it to mind by my own volition. The "deep unconscious" of contemporary cognitive science is also quite unlike Freud's "dynamic unconscious" (repressed memories, desires, etc.), although the theory--and controversies--of the latter need not detain us here. But at least repressed mental states could potentially come to the surface via therapy. The unconscious mental states posited by contemporary cognitive science are much more hermetically sealed.

I will use mental-logic theories (MLT) to anchor my discussion, but the arguments I will be making will apply to other computational accounts of reasoning, such as mental-model theory. I believe that it might be possible to adapt these arguments in a way that will make them applicable to any theory that postulates unconscious computation, including theories of low-level peripheral cognition such as perception and language. But in what follows I will only be concerned with computational theories of reasoning. For simplicity, I will restrict attention to propositional logic, and specifically to what is often called the "logical judgment" problem, whereby a small number of fairly simple premises are given (often just one premise), along with a putative conclusion, and the problem is to determine whether the conclusion follows deductively from the premises.

Alice is a college sophomore without any training in formal logic, although perhaps she has a meager background in algorithms (e.g., she might know what an algorithm is, and have a vague notion of what loops and conditional branches are for). According to mental-logic doctrine, Alice is equipped with a module for reasoning in propositional logic that consists of:

(1) a number of inference schemas, such as modus ponens; and

(2) a control procedure, which, presented with a reasoning problem, regulates the selection of which inference rules to apply, when to backtrack, and so on.

The procedure always terminates and can result in an affirmative, a negative, or an inconclusive ("can't tell") answer. (In the context of the logical judgment problem, these two components are sometimes called the "operations" and the "executive," respectively.) Now let  $L$  be Alice's logic for "logical judgment" in propositional logic, and let  $R$  be the associated procedure. And let  $P$  be a simple propositional reasoning problem. Presumably, if we presented Alice with  $P$ , her mental logic would kick in,  $R$  would operate for a finite period of time, and before long an answer would emerge.

The contents of both  $L$  and  $R$  are in thinkable form, and indeed are eminently learnable.  $L$  presumably contains such straightforward inference rules as the contrapositive, and  $R$  contains a small number of simple instructions such as conditional branching and looping. It is quite conceivable, therefore, that Alice can be taught the specific rules of  $L$  and the algorithm  $R$ , and can voluntarily and consciously follow  $R$ . This does not have to be deliberate, in that I am not assuming that  $L$  and  $R$  are taught to Alice as the very mental logic that her own mind contains for propositional logical judgment. They could be taught to her fortuitously, as part of a random teaching assignment by a teacher, or by some instructor as part of a cognitive science experiment, and it could just so happen, by accident, that what she is taught is in fact identical to her "mental logic," although Alice herself is entirely unaware of this. In fact Alice might not even be aware that she has such a logic at all.

Now suppose that after a short crash course on  $L$  and  $R$ , Alice is presented with problem  $P$  and goes to work consciously applying  $R$ , while, unconsciously and unbeknownst to her, she is applying the very same procedure at the same time. The exact same process unfolds in two duplicate and concurrent threads, tracing two sequences of intentional states, which I will write as  $s_1, \dots, s_n$  for the conscious process and  $s'_1, \dots, s'_n$  for the unconscious one. We might allow—as is surely logically possible, though improbable—that the concurrency is exact, and that the two threads proceed in perfect lockstep. I claim that  $s_i$  and  $s'_i$  are identical intentional state tokens for each  $i = 1, \dots, n$ . We might say that two intentional states are type-identical if they have the same mode and the same content (propositional or otherwise), so, for instance, your belief that Obama is the president of the USA is type-identical to my belief that Obama is the president of the USA because both the psychological mode (belief) and the content (that Obama is the president of the USA) are identical. What are reasonable identity criteria for intentional state tokens? Two intentional state tokens of one and the same person are identical if they have the same mode, the same content, the same causes, and sufficient temporal proximity.

In the present scenario, all these conditions obtain. Content and mode are identical by virtue of the fact that the logic and the algorithm on both levels are identical, and the causes are also the same in both cases—the execution of that particular algorithm on that particular input. Remember that according to the standard computational theory of the mind, the algorithms that are postulated by various cognitive scientists involve intrinsic intentionality (i.e., they are not observer-relative), and are causally efficacious. That is, a person's cognitive activity and concomitant intentional states are

the way they are because he or she is running the algorithm in question. So in both cases, it is the deployment of the same algorithm on the same input that is causing the states. Of course in this version of the thought experiment we actually have more than that. We also have complete temporal overlap. So, for any  $i$ , both  $s_i$  and  $s_i'$  are occurring at the exact same time, in the same mind, with the exact same contents, and the exact same causes and effects. Therefore, the states are identical. But this is a contradiction, because we are now led to admit that one and the same intentional state is simultaneously occurring both consciously and unconsciously. I regard the contradiction as a reductio of the hypothesis that the process  $s_1', \dots, s_n'$  is occurring unconsciously; that the process  $s_1, \dots, s_n$  is consciously occurring is, of course, beyond doubt. I conclude that there are no such unconscious intentional states. The only intrinsic intentional states and computational processes that actually take place are the conscious ones.

## OUTLINING A COMPUTATIONALLY PLAUSIBLE APPROACH TO MENTAL STATE ASCRIPTION

WILL BRIDEWELL  
*Center for Biomedical Informatics Research*  
*Stanford University, Stanford, CA USA*

AND

Alistair Isaac  
*Department of Philosophy*  
*University of Michigan, Ann Arbor, MI USA*

AND

Pat Langley  
*Computer Science and Engineering*  
*Arizona State University, Tempe AZ USA*

### 1. Extended Abstract

No one would debate that social cognition is a key characteristic of human-level intelligence. However, within the artificial intelligence literature, we find no system that carries out more than a rudimentary level of social interaction. Previous theoretical work on social information processing usually treats agents as input–output systems that lack internal representations of each other (e.g., multiagent systems) or develops formalisms unsuitable for practical implementation (e.g., undecidable epistemic logics). To move forward, new strategies for modeling interaction need to tractably support reasoning about the mental states of oneself and others. Here, we present steps toward such a model that we hope will address the need for a computationally plausible approach and will eventually lead to a system that can engage in complex dialog with others.

An agent’s mental space is partitioned into models of agents. One of these is the model of self, which serves as the default source-of-truth when reasoning about the world. From a computational perspective, we find it useful to separate different modalities of mentality into different regions. For instance, inside of its self model, an agent will have a structure that stores beliefs about the state of the world, one that stores goals that indicate desired future states of the world, and one that stores intentions which are actions that manifest the goals. Since goals and intentions in this representation refer to mental states of which the agent is aware, we loosely use those terms as shorthand for the agent’s beliefs about its goals and beliefs about its intentions. Taking this view, the primitive mental object is the belief.

Continuing with the computational perspective, we represent a belief as a data structure that contains a literal representing its content and other contextual features necessary to guide reasoning. These features include temporal aspects analogous to valid time and transactional time in a database. That is, the literals in the belief may be associated with the period of time for which they were true (e.g., Yesterday, Jeff ate lunch between 11 and 12) and the period of time during which they were held (e.g., I believed that Chris was a man until I met her), both of which may overlap. Asserting a belief as a goal or an intention involves placing it in the appropriate mental partition and does not require a corresponding change in representation.

In addition to beliefs, which are stored within agent models, we represent relationships among those models. The principal agent model (i.e., the model of the self) connects to internalized models of other agents. These models are accessible through a *believes* relation. For example, consider a technical support agent conversing with a customer. During the exchange, the support agent may reason about whether the customer believes that his computer is plugged in. Trivially, we might represent this statement as (belief Customer (plugged-in computer)), which tells the system implementing the agent to look in the beliefs of the Customer model believed by the principal agent. Continuing, the agent may have a goal (goal (belief Customer (not (plugged-in computer))))). This goal would appear in a second Customer model that is connected to the agent's goal space instead of its belief space. Notably the *goal*, *intention*, and *belief* operators are not modal operators. For our purposes, they index mental spaces that contain sets of beliefs.

Importantly, knowledge is stored only when necessary. The principal agent's default assumption is that other agents' beliefs are in accord with its own. If the principal agent has no reason to believe that another agent is in disagreement, then that agent's model will be empty. In the previous example, if the agent believes (plugged-in computer) and (believes Customer (plugged-in computer)), the actual belief will only appear in the principal agent's model. The other models inherit the beliefs of their parents via default reasoning unless a specific belief is overridden by a locally stored, incompatible one such as (not (plugged-in computer)). As a rough approximation, we assume that all agents share the same inference mechanisms and long-term knowledge (e.g., rules) and do not attempt to represent differences in cognitive ability or domain knowledge.

With this basic framework in mind, there are six challenges that must be addressed to implement a functioning system. Here we present these along with our proposed solutions for two of the most compelling ones.

1. *When are new agent models introduced?*
2. *When are agents linked to each other?*
3. *How are agents traversed to unpack a nested statement?*
4. *What is taken as common ground?*
5. *How are beliefs ascribed to nested agents?*
6. *How does one agent reason about another?*

Addressing the first challenge, the most apparent situation is when a new agent joins a conversation. If individuals discuss an absent agent, one may treat that agent as either a simple object or an agent to whom one may ascribe beliefs. To illustrate, suppose Tom tells the principal agent, "Harry likes pudding." That would correspond to some belief either in the principal model or the Tom model that resembles (likes Harry pudding). If

instead, Tom said, “Harry said that he likes pudding,” we would need to create a model of Harry, that would let us store (believes Harry (likes Harry pudding)). Where the belief resides depends on the mental state of the other agents and how their models are connected.

Answering the sixth challenge, we recall that all agents are assumed to use the same inference system and domain knowledge as the principal agent. Typically this mechanism “resides” in that agent’s model. However, one can shift perspective by moving the seat of the inference system to another agent model. In this sense, there is a clear relationship to simulation theory, but the domain knowledge may include rules that encode how agents reason about each other much like the theory-theory. As a result, we can integrate ideas from both camps to help reach our operational goal of intelligent systems that can collaborate and engage with people in realistic dialogs.

### **Acknowledgements**

Will Bridewell and Pat Langley are funded by the Office of Naval Research under Contract No. ONR-N00014-09-1-1029. Alistair Isaac is funded by a postdoctoral fellowship from the McDonnell Foundation Research Consortium on Causal Learning.

## **AGENCY: ON MACHINES THAT MENTALIZE**

MARCELLO GUARINI

*University of Windsor*

*401 Sunset, Windsor, ON, Canada N9B 394*

### **1. Agency, Responsibility, and Mentalizing**

The ability of human beings to attribute mental states has been variously referred to as “mindreading” and “mentalizing.” The purpose of this paper will be to examine the relationship between agency and mentalizing.

Two dimensions of agency will be discussed. The first is the ability of a human or machine to take responsibility for his/her/its actions and thoughts – a first person ability. The second is the ability to hold others responsible – a third person ability. Both of these activities are important for various forms of social interaction, and they would not be possible without mentalizing. It will be shown that various mindreading abilities – such as tracking perception, desire, the source of belief, and false belief – are central to the notion of agency in ethical, epistemic, and legal contexts. This has implications not only for how we understand human agency, but for how we understand the agency of future machines.

### **2. Conditions of Agency**

Agency comes in degrees: we might expect an average five year old human child to take responsibility for some things, and an average 15 year old to take responsibility for still further things, and an average 25 year old to take responsibility for still further things. We should expect variations in the capacities of machines as well. The focus of this work will be the kinds of mentalizing tasks that average five year olds excel at, and the contribution they make to understanding agency. A framework will be provided for understanding the conditions of agency. Distinctions will be made between the generative conditions of agency (what it takes to bring agency into existence), the maintenance conditions of agency (what is required to keep agency in existence), and the regenerative conditions of agency (what is required to repair or restore agency if it is impaired). It will be argued that sustaining various mentalizing abilities are among the maintenance conditions of agency.

#### **2.1. AN EXAMPLE**

Let us consider the capacity to attribute false beliefs, something most 5 year olds possess. Some children are allowed to view a Smarties box that has candy (Nichols and

Stich, 2003, p.90). One of the children is asked to leave the room, and the remaining children witness the candy being replaced with pencils. The absent child is brought back into the room. When asked what the temporarily absent child believes is in the box, most three year olds say “pencils.” This is a third person failure to attribute a false belief. Tasks such as these can be failed in the first person as well: young children often fail to attribute false beliefs to themselves. There is an important connection between agency and the ability to attribute false beliefs. The ability to take responsibility involves, among other things, the ability to grasp that I have or had a false or incorrect view. Without the ability to attribute error to oneself, it is difficult to see how one could in some well developed sense take responsibility for it. Moreover, holding another responsible could well involve, among other things, attributing a false belief to that other individual. Agent  $A_1$  may challenge  $A_2$  to revise his, her, or its view on some matter on the grounds that the view is false.  $A_1$  needs to be able to attribute a false belief to  $A_2$  for this to happen.

## 2.2. LEVELS AND CONDITIONS OF AGENCY

There is some recent research that uses an attentional (as opposed to linguistic) paradigm to argue that children engage in some sort of false belief recognition well before language is developed (Goldman, 2006, pp. 76-77). This is startling and interesting work, but whatever these very early abilities amount to, it will be argued that they are insufficient for understanding what is required in advanced forms of taking responsibility or holding others responsible. They will, however, play an important role in understanding the generative conditions of human agency. Success in these very early attentional tasks appear to be important precursors to the linguistic abilities required for advanced forms of agency. Supporting what is needed for these attentional abilities might also be among the maintenance conditions of simpler forms of agency.

A discussion of the conditions of agency can be usefully augmented with the well worn three level approach to explanation common in cognitive science – intentional, algorithmic or mathematical, and implementational. We can examine the conditions of agency at each of these levels. For example, at the intentional level, we can intentionally specify what sorts of abilities have to be kept in place or maintained for advanced agency to exist – much of this may be the same for humans and machines. However, at the algorithmic/mathematical and implementational levels, there may be important differences in specifying how agency is maintained.

## 3. Significance

At some point, we expect our children to start taking responsibility for their behaviour and engage in self-correcting behaviour made possible by false belief attribution and other mindreading abilities. Among other things, this creates various epistemic, moral, and other efficiencies – individuals that can monitor and correct their own thoughts and behaviours do not require constant correction from others, which frees these agents to pursue further tasks. One of the driving forces behind the development of machine agency will no doubt be the desire for these sorts of efficiencies. It will be shown that other mindreading tasks (over and above false belief attribution) play a role in first and third person dimensions of agency.

## References

- Nichols, S. & Stich, S.P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.
- Goldman, A.I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

## TOWARD A TESTBED FOR MODELING THE KNOWLEDGE, GOALS AND MENTAL STATES OF OTHERS

SERGEI NIRENBURG

*University of Maryland Baltimore County  
Baltimore, MD, 21250 USA*

**Abstract.** The paper introduces a computational environment that facilitates development and experimentation with intelligent agents in the OntoAgent cognitive architecture. The agents pursue goal- and plan-oriented reasoning, are capable of communicating in natural language and build mental models of other agents.

Decision-making is a core capability of intelligent agents – both human and artificial ones. Making optimal decisions with limited resources is a very difficult task both for people and for machines. Helping people to make decisions is an important scientific, societal and technological goal.

Classical decision theory presupposes an idealized decision-making agent that possesses *all the knowledge* necessary (or desired) for making a decision, operates with *optimum decision procedures* and is fully *rational* in terms of the rational choice theory. Within this theory rationality of an individual decision is estimated in terms of what von Neumann and Morgenstern (1944) called expected utility, the cost effectiveness of the means to achieve a specific goal. In other words, rational behavior for an individual maximizes benefits and minimizes costs of a choice.

However, in real life few people make decisions under conditions of complete knowledge, maximum efficiency and rationality. Thus, Simon (1955) introduced the concept of bounded rationality that removes the constraint of having complete knowledge and the best algorithm by switching from seeking an optimal decision to accepting a *satisficing* decision (roughly, making do with the first decision for which utility exceeds costs even though there may be any number of better decisions available). A number of proposals concentrated on the selection of parameters (features) on the basis of which choices are made. Thus, the prospect theory of Tversky and Kahneman (1974) and its descendants, such as cumulative prospect theory, augment the inventory of decision parameters for a decision (utility) function by stressing psychological influences on decision-making, such as risk aversion and “reference” utility meaning utility relative to perceived utility for others.

In order to incorporate the latter, an intelligent agent  $A_0$  must be able to model the mental states of other agents,  $A_1, \dots, A_n$ . At the intuitive level, we understand mental states as including, at a minimum, ontological knowledge of concept types as well as knowledge of concept instances, the agent’s goals, preferences, personality traits, etc. The concept of ‘belief,’ often used in conjunction with modeling agents we interpret as

(possibly, error-ridden) knowledge that agent  $A_0$  has about other agents it knows. (We are aware that the knowledge  $A_0$  has about itself may also be less than accurate.)

In our work on modeling intelligent agents we stress the importance of extending the inventory of an agent's decision-making parameters (but only if effective procedures for determining their values can be developed). Thus, it is correct to state that understanding speaker's goals is important in making a decision about how to react to a speech act. But in practice more specific knowledge is needed – for example, when a doctor asks a patient about the latter's family, the patient must judge whether the speaker's goal is professional (having the patient's condition diagnosed) or social (making small talk) or – and this is an even more complex reckoning – whether it is a social goal put in service of the professional one (aiming at establishing a rapport with a patient so as to develop trust and ensure cooperation – better-quality responses to questions and requests).

In this talk I will describe a computational environment that facilitates development and experimentation with agents that strive to make use of mental models of others as a prerequisite for making appropriate decisions with respect to the agent's own behavior. This capability is one of several core requirements of our cognitive architecture, OntoAgent. In addition to modeling ontological knowledge about the outside world and knowledge about remembered instances of ontological concepts (including other agents, viewed as instances of the ontological concept HUMAN), OntoSem agents:

- are designed to operate in a hybrid network of human and artificial agents;
- emulate human information processing capabilities by modeling conscious perception and action;
- communicate with people using natural language;
- can incorporate a physiological model, making them what we call “double agents” with simulated bodies as well as simulated minds;
- can be endowed with personality traits, preferences and psychological states that influence their perceived or subconscious decision-making preferences;
- rely on knowledge resources and processors that are broad-coverage rather than geared at a particular application, which simplifies porting agents to new domains and applications;
- stress the importance of memory of event, state and object instances to complement its ontological knowledge of event, state and object types.

What makes modeling such multi-faceted agents feasible is that all aspects of agent functioning are supported by the same knowledge substrate encoded in a single metalanguage. The OntoAgent testbed has been implemented in the medical domain and supports two agent environments:

- Maryland Virtual Patient (MVP, McShane et al. 2009) modeling a patient, a trainee MD and a tutor in the process of learning medical diagnostics and treatment; and
- CLinician's ADdvisor (CLAD, Nirenburg et al. 2011) modeling a patient, an MD and a clinician's advisor and intended to assist practicing clinicians by reducing their cognitive load.

The talk will include a demonstration of the above environments and a discussion of the ways of modeling mental states of other agents.

## References

- McShane, M., S. Nirenburg, B. Jarrell, S. Beale, G. Fantry (2009). Maryland Virtual Patient: A Knowledge-Based, Language-Enabled Simulation and Training System. Proceedings of International Conference on Virtual Patients, Krakow, Poland, June 5-6.
- Neuman, J. von and O. Morgenstern (1944). *Theory of Games and Economic Behaviour*. Princeton: Princeton University Press.
- Nirenburg, Sergei, Marjorie McShane, Stephen Beale, Bruce Jarrell and George Fantry (2011). Intelligent agents in support of clinical medicine. Proceedings of MMVR18, Newport Beach, CA, February 9–12.
- Simon, H.A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 69: 99–118, 1955.
- Tversky, Amos, & Kahneman, Daniel (1974). Judgment under uncertainty: Heuristics and Biases. *Science*, 185: 1124-1131.

## ARCHITECTURAL STEPS TOWARDS SELF-AWARE ROBOTS

MATTHIAS SCHEUTZ

*Tufts University*

*161 College Ave., Medford MA 02155*

**Abstract.** Philosophical debates about qualia, perspectivalness, “what it is like” experiences and related topics are vastly disconnected from “architecture talk” in AI and cognitive science which is required for understanding minds and designing artificial agents. While philosophy can thus not help AI in designing conscious agents, I argue that AI and robotics cannot only help philosophy, but may even be required for solving some of the puzzling questions in the philosophy of consciousness. Specifically, I will claim that there is no such thing as a necessarily private experience (neither phenomenal, nor introspective, nor any other) using as an example robotic architectures whose instances “know” what it is like to be another robotic architecture instance.

Start with two basic hopefully non-controversial notions, those of *awareness* and *self-awareness*, define them for agent architectures and then show how we can say that a robot is *aware* or *self-aware* in a given context. Following Chalmers' (1996) notion of *awareness* and Block's (1995) notion of *access consciousness*, call a state *S* of an agent architecture *A* an “awareness state” if *S* contains information *about something* (entity, state, event, etc.) that the agent (instantiating *A*) can use to make decisions, guide its behavior and/or give verbal reports. Specifically, an agent is “aware of *X*” if it is in an awareness state that in some way represents or encodes *X*. An agent is “self-aware” if it is aware of itself, i.e., if it is in an awareness state that represents or encodes (parts of) the agent itself. *S* will typically be a *complex state* that consists of substates reflecting the states of various functional components in the architecture *A*. For example, if *S* is the state of “being aware of a red box”, then this state will roughly require perceptual states representing the box and some of its properties including its redness, in addition to states that use some of these representations in order to form other representations and/or behaviors.

To make all of this more precise, I will briefly introduce some relevant parts of our robotic DIARC architecture that we have been developing over the last decade or so in my lab (Scheutz et al 2007). What is nice about robotic architectures (or any form of agent architecture, including cognitive architectures for that matter) is that one can look inside. I.e., one can take a look at the blueprint and follow the information flow along connections between functional components. One can trace processing routes and look at component states. And one can make statements about possible and impossible processes in a system that instantiates the architecture.

DIARC consists of various functional modules: on the perception side, there are modules for vision processing, sound processing (including sound localization and speech recognition), laser distance data processing, and processing of various internal proprioceptive sensors. For most sensory modalities, there are also short and long-term memories, e.g., a long-term memory for visual objects and a short-term memory for storing the recognized objects the agent currently sees. On the action side, there are modules for moving the robot body through the environment, for making arm and head movements, and for making facial expressions, among others. Internal modules consist of various short and long-term memories together with processes that operate on those memories, including skill memories, factual and episodic memories, a lexicon with syntactic and semantic annotations in addition to word forms, and a task memory. Moreover, there are components for managing the agent's goal, for scheduling actions in parallel, for processing spoken natural language, for task planning, and for reasoning (for more details, see Scheutz et al. 2007).

Now consider a robot running DIARC that is asked whether it sees a red box and assume that the robot has a goal to answer questions. Upon hearing the spoken utterance, the speech recognizer generates word tokens from it, which are then syntactically and semantically analyzed, resulting in an internal logical representation of the meaning. The robot recognizes that the utterance was a question that required it to perform an internal lookup action in its visual short term memory (VSTM), namely to check whether VSTM contains an object representation of a red box. Note that the robot only needs to perform a lookup action in its VSTM, because VSTM is automatically updated based on what the object recognition algorithm detects in the image coming from the camera at a rate of 30Hz. In particular, various vision processing algorithms are performed on each image frame attempting to segment colored regions, detect object boundaries, recognize objects and determine their properties. These processes result in the generation of representations of the recognized objects in VSTM, which are matched against existing representations so that object identities can be tracking over short periods of time. If the agent has an object representation of a red box in VSTM, then the representation is retrieved and bound to the expression "red box". The binding confirms the resolution of the reference and triggers a variety of additional bindings (including the binding of various discourse variables such as "last mentioned object" and "last mentioned noun" in linguistic short-term memory). It also triggers the generation of an answer to confirm that the robot is seeing a red box, which the robot then pronounces. In addition, the generated answer gets stored in linguistic short-term memory and, depending on other factors, the whole event "you asked whether I saw a red box, and I did see one" might get stored in episodic memory (indexed by time, object type, interaction type, and others).

From the above description, it is clear that the robot went through several awareness states including self-awareness states as part of answering the question: the robot is aware of the question when it is in a state where it checks for the object asked for in the question; if there is such an object, the robot becomes aware of the object as well as of the object's properties (in particular, its color), and the robot is aware of the answer it gave. Moreover, the robot is aware of itself having been asked the question and of having given the answer, which is a self-awareness state.

I will then use the above architecture to demonstrate during my presentation what it is like for the robot to have a color experience and use this result to address some questions about phenomenal and private experience in philosophy. In particular, I will argue that robots *can know* what it is like to have another robot's experience.

## References

- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-247.
- Chalmers, D. J. (1996a). *The conscious mind: In search of a fundamental theory*. New York, NY, USA: Oxford University Press.
- Scheutz, M., Schermerhorn, P., Kramer, J. and Anderson, D. (2007). "First Steps toward Natural Human-Like HRI". *Autonomous Robots*, 22, 4, 411-423.

## LOGIC-BASED SIMULATIONS OF MIRROR TESTING FOR SELF-CONSCIOUSNESS

NAVEEN SUNDAR

AND

SELMER BRINGSJORD

**Abstract.** We present a formal logic-based analysis of the mirror test for self-consciousness. Based on this formalization, a computational simulation of a mirror-failing dog, a mirror-passing chimp, and a mirror-passing human will be presented. The simulation will consist in the automatic machine-found disproof in the case of the canine, and proofs in the other two cases. These simulations will be based on an axiomatization of the perceptual and doxastic details assumed to be in/operative in these three cases by those embracing the view that chimps and humans are self-conscious, while dogs aren't.

### 1. The Mirror Test

In accordance with a now-familiar recipe **R** in the annals of the study of “self-consciousness,” anesthetize<sup>23</sup> a creature  $c$ ; while it's under, paint, say, a red (odorless, hypo-allergenic) splotch upon its forehead, thus making it true that  $c$  has property  $R$  ( $= Rc$ ); when awake, place  $c$  in front of a mirror ( $Mc$ ); observe the creature's behavior  $b$  to see if it for example includes the attempt to remove the splotch ( $Rcb$  or  $\neg Rcb$ ); if it does/doesn't, issue a pronouncement about such questions as whether or not it's self-conscious (or self-aware, etc.; i.e., as to whether or not  $Sc$ ).

Descriptions of the following of **R** are innumerable in the literature.<sup>24</sup> But what is the logic of this recipe? Despite decades of writing about the value of the recipe, we can find no rigorous account of it, nor of followings of it in connection with certain classes of creatures. Therefore, we can't find rigorous computational simulations of such followings, and we certainly can't find proofs that for given creatures they are known to either have or lack self-consciousness, depending upon whether or not they pass the mirror test. Work underway by us is designed to provide these missing things, and we propose to report on this work at IACAP 2011, and show demonstrations.

---

23 Or perhaps do it while the creature is sleeping soundly.

24 For a compendium of such followings, accompanied by the colorful proposal that self-awareness can be neuro-localized in the right hemisphere, see Keenan, J., Gallup, G. and Falk, D. *The Face in the Mirror* (Ecco: New York, NY).

## 2. Toward a Formal Analysis of the Mirror Test

Let's assume a standard extensional multi-sorted logic in which creatures are partitioned in customary ways. (Please note that the empirical, informal literature, as a matter of brute fact, makes not even a nod in the intensional direction, and is naturally formalized via extensional frameworks.) Specifically, the class of dogs will be denoted by 'D,' chimps by 'C,' and humans by 'H.' Then, the following three propositions have apparently been affirmed in the literature.

1.  $\Box c \Box D [(Rc \Box Mc \Box Rcb) \rightarrow Sc]$  • This is taken to be true, in a nutshell, because if dogs had behaved as chimps usually do, canines would have presumably been admitted into the "self-aware" club.
2.  $\Box c \Box C [(Rc \Box Mc \Box \neg Rcb) \rightarrow \neg Sc]$  • This is taken to be true, in short, because if chimps had behaved as dogs do, chimps would have presumably have been kept out of the "self-aware" club.
3.  $\Box c \Box H [(Rc \Box Mc \Box \neg Rcb) \rightarrow \neg Sc]$  • This is taken to be true, in a nutshell, because humans provide the "anchor point" on the issue at hand.

Unfortunately, none of these propositions are true. A dog pre-trained to paw its forehead when seeing a dog provides a counter-example to 1., since no participant in the debate herein considered accepts that such training ensures self-consciousness.<sup>25</sup> A chimp pre-trained to leave splotches intact constitutes a counter-example to 2., since no participant accepts that such training guarantees the absence of self-consciousness. And a human inclined to ignore splotches overthrows proposition 3.

Of course, these problems are just the tip of the iceberg. The trio is of course incomplete, since from them one cannot for instance deduce that dogs aren't self-conscious, whereas chimps and humans are. One might think that this is addressed by adding more formulae<sup>26</sup>, but since the conditional used here is the material conditional, this trio can't possibly be heading in the right direction, as is easily seen. Assume that a variant of 2., 2', is to enable deduction that some real-life chimp, Charlie,  $c$ , is in fact self-conscious. How could this deduction go through? It could only work if the relevant antecedents in 2' were satisfied. For example, the following holds.

$$\{2'\} \Box \{Rc \Box \Box Mc \Box \Box Rc \Box b\} \Box Sc \Box$$

But for Charlie, and nearly every single chimp who ever lived or will ever live, there will never be a red splotch and a mirror in his life. And yet clearly those in favor of ascribing self-consciousness to chimps will want to make the ascription to Charlie and his friends. More specifically, those in favor of the ascription presumably hold that were it the case that Charlie was given the mirror test, he would pass. This indicates that some intensional logic is required; specifically, a conditional logic able to handle subjunctive conditionals is needed.

---

25 Of course, someone might deny that such behaviour expresses an intention to remove a splotch, but that would be entirely ad hoc. Trainers after all routinely train dogs to form goals and seek their satisfaction when they observe the relevant triggers. Relevant here is the Keenan-et-al.-recounted story of behaviourists who claimed that pigeons were to be classified with chimps in the running of R. It turned out that the pigeons had been pre-trained in ways that contaminated the experimentation in question.

26 E.g.,  $\Box c \Box D [(Rc \Box Mc \Box \neg Rcb) \rightarrow \neg Sc]$ .

Note that the fact that  $2'$  might never be satisfied for a particular chimp is not the fault of our chosen formulation, since that formulation is a direct symbolization of what is said in the literature (which has of course been written for the most part by informalists). One way to understand what ought to be claimed in the informal literature is that a subjunctive conditional be employed: for example, if in all nearby “possible worlds” in which  $R_c$  and  $M_c$  are true,  $R_{cb}$  is true, then  $S_c$  is true in the actual world. But of course this sort of thing is the point, since no one has yet worked out the details in this direction, and to credit this direction to anyone in the empirical prior work is so charitable as to border on absurdity. And of course the devil is in the details: The formal calculi we use include an explicit rejection of a possible-worlds semantics for anything doxastic.

Our modeling of mirror testing has obvious connections to key distinctions recently made by Clowes and Seth (2008). In their terms, our research is without question “weak” in nature, since we don’t claim that our mirror-passing agents, however formal and fine-grained the underlying modeling may be, literally are conscious. In addition, while elsewhere (Bringsjord 2007) one of us has expressed skepticism about Aleksander’s axiomatic approach, discussed by C&S, our approach is certainly axiomatic. However, the calculi upon which this approach rests are more expressive than those used by Aleksander (allowing, e.g., for intensional operators), and are oriented toward proof theory and automated proof finding and checking.

Finally, related prior work in simulating the mirror test can be found in Takeno’s work on mirror image discrimination. This work provides some evidence that at least the rather informal robotics side of the act of a simple agent’s recognizing its mirror image is feasible. We will of course contrast our work with that of Takeno et al.

## References

- Bringsjord, S. (2007). Offer: One Billion Dollars for a Conscious Robot. If You’re Honest, You Must Decline. *Journal of Consciousness Studies*, 14(7), 28–43.
- Clowes, R.W. & Seth, A.K. (2008). Axioms, Properties and Criteria: Roles for Synthesis in the Science of Consciousness. *Artificial Intelligence in Medicine*, 44(2), 91-104.
- Takeno, J. & Inaba, K. & Suzuki, T. (2005). Experiments and Examination of Mirror Image Cognition Using a Small Robot. Proceedings of CIRA 2005: *IEEE International Symposium on Computational Intelligence in Robotics and Automation*. Espoo, Finland, 2005.

## List of Authors in Alphabetic Order

Aas, Katja Franko	25
Alhutter, Doris	252
Anokhina, Margaryta	119
Arkin, Ronald	122 & 317
Arkoudas, Konstantine	320
Asai, Ryoko	287
Asaro, Peter	179
Backhaus, Patrick	290
Barker, Steve	255
Baumgaertner, Bert	206
Beavers, Anthony F.	23
Beinsteiner, Andreas	301
Belfer, Israel	209
Bello, Paul et alia	125
Bengez, Rainhard Z.	33
Blanco, Javier O.	34
Bod, Rens	216
Boltuc, Peter	38
Breems, Nick	213
Bridewell, Will	323
Briggs, Gordon	128
Bringsjord, Selmer	335
Buchanan, Elizabeth	30
Buckner, Cameron	29
Bynum, Terrell Ward	26
Calabretto, Sylvie	242
Casacuberta, David	143
Chokvasin, Theptawee	41
Coeckelbergh, Mark	258
Cohen, Paul	95
Compagna, Diego	262
Crutzen, C.K.M.	156
Danka, Istvan	264
Dasch, Thomas	181
De Gooijer, Thijmen	293
Desclés, Jean-Pierre	220
Dodig-Crnkovic, Gordana	119 & 262 & 290
Douglas, Keith	184

Duran, Juan M.	44
Ekbia, Hamid R.	247 & 269
Ess, Charles	30
Franchette, Florent	47
Franchi, Stefano	223
Funcke, Alexander	83
Ganascia, Jean-Gabriel	304
Geier, Fabian	50
Giardino, Valeria	87
Guarini, Marcello	228
Guarini, Marcello	326
Hagengruber, Ruth	131
Halpin, Harry	57
Heersmink, Richard	91
Heimo, Olli I.	133
Hempel, Leon	159
Hewlett, David	95
Hongladarom, Sonja	298
Hromada, Daniel D.	186
Janlert, Lars-Erik	98
Kavathatzopoulos, Iordanis	137
Kimppa, Kai K.	133
Kitto, Kirsty	101
Laaksoharju, Mikael	137
Macnish, Kevin	163
Mauger, Jeremy	30
McKinley, Steve	231 & 234
Menant, Christophe	105
Meyer, Steven	53
Molyneux, Bernard	140
Monin, Alexandre	57
Najar, Anis	307
Nicolaidis, Michael	238
Nirenburg, Sergej	329

Othmer, Julius	166
Pagano, Miguel	60
Portier, Pierre-Edouard	242
Quiroz, Francisco Hernandez	108
Reynolds, Carson	310
Riss, Uwe	64
Ropolyi, Laszlo	273
Scheutz, Matthias	332
Schroeder, Marcin	111
Simon, Judith	276
Sinclair, Nathan	68
Smith, Lindsay	71
Solodovnik, Iryna	75
Soraker, Johnny Hartz	190
Strauss, Stefan	313
Sullins, John P	28
Sundar, Naveen	335
Taddeo, Mariarosa	168
Thürmel, Sabine	78
Tonkens, Ryan	194
Turner, Raymond	81
Vakarelov, Orlin	115
Vallor, Shannon	197
Vallverdu, Jordi	143
Veale, Richard	147
Vehlken, Sebastian	279
Waser, Mark R.	152
Weber, Jutta	172
Weich, Andreas	166
Wong, Pak-Hang	201
York, William W.	247
Zambak, Aziz	282
Zhang, Guo	269