

Cognitive Systems Monographs

Volume 42

Series Editors

Rüdiger Dillmann, University of Karlsruhe, Karlsruhe, Germany

Yoshihiko Nakamura, Department of Mechano-Informatics, Tokyo University, Tokyo, Japan

Stefan Schaal, University of Southern California, Los Angeles, CA, USA

David Vernon, University of Skövde, Skövde, Sweden

Advisory Editors

Heinrich H. Bülthoff, MPI for Biological Cybernetics, Tübingen, Germany

Masayuki Inaba, University of Tokyo, Tokyo, Japan

J.A. Scott Kelso, Florida Atlantic University, Boca Raton, FL, USA

Oussama Khatib, Stanford University, Stanford, CA, USA

Yasuo Kuniyoshi, The University of Tokyo, Tokyo, Japan

Hiroshi G. Okuno, Kyoto University, Kyoto, Japan

Helge Ritter, University of Bielefeld, Bielefeld, Germany

Giulio Sandini, University of Genova, Genova, Italy

Bruno Siciliano, University of Naples, Napoli, Italy

Mark Steedman, University of Edinburgh, Edinburgh, UK

Atsuo Takanishi, Waseda University, Tokyo, Japan

The Cognitive Systems Monographs (COSMOS) publish new developments and advances in the fields of cognitive systems research, rapidly and informally but with a high quality. The intent is to bridge cognitive brain science and biology with engineering disciplines. It covers all the technical contents, applications, and multidisciplinary aspects of cognitive systems, such as Bionics, System Analysis, System Modelling, System Design, Human Motion Understanding, Human Activity Understanding, Learning of Behaviour, Man-Machine Interaction, Smart and Cognitive Environments, Human and Computer Vision, Neuroinformatics, Humanoids, Biologically motivated systems and artefacts, Autonomous Systems, Linguistics, Sports Engineering, Computational Intelligence, Biosignal Processing, or Cognitive Materials—as well as the methodologies behind them. Within the scope of the series are monographs, lecture notes, selected contributions from specialized conferences and workshops, as well as selected Ph.D. theses.

Indexed by SCOPUS, DBLP, zbMATH, SCImago.

More information about this series at <http://www.springer.com/series/8354>

Joachim Diederich

The Psychology of Artificial Superintelligence

Joachim Diederich
The University of Queensland
School of Information Technology and Electrical
Engineering
St Lucia, QLD, Australia

ISSN 1867-4925

ISSN 1867-4933 (electronic)

Cognitive Systems Monographs

ISBN 978-3-030-71841-1

ISBN 978-3-030-71842-8 (eBook)

<https://doi.org/10.1007/978-3-030-71842-8>

© Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

For Susan

Preface

This book explores the psychological impact of advanced forms of artificial intelligence. There are psychological consequences for the well-being of individuals as well as a significant impact on societies. Human work will continue to be transformed and will possibly be eliminated in the not-so-distant future. Interfaces that directly connect the brain with the internet will have an impact on how we think and communicate. The decisions and actions of an advanced form of artificial intelligence will be more and more difficult to understand, and hence, better forms of explanation for artificial intelligence are required. The technology is increasingly being used to manage significant parts of society, e.g. by use of social credit systems, with consequences for the entire population. Finally, advancements in military AI may include autonomous killing machines that can spread fear and terror. All these developments are happening as we speak and represent significant challenges to human psychological well-being.

Clearly, there are advancements in the medical sector through improved artificial intelligence, and human lives may be enriched in other areas as well. But there needs to be an informed discussion about the risks and challenges of such advancements. There needs to be information so that individuals can make a decision on the use of the technology or the exposure to new developments.

The core message of this book is that advanced forms of artificial intelligence will have an impact on everybody: The developers and users of AI systems as well as individuals who have no direct contact with this form of technology. This is due to the *soliciting* nature of artificial intelligence. The *universal solicitation* of the technology is the challenge. A solicitation that can become a demand.

Machines “want” to be used because they directly relate to human needs. This goes beyond the concept of affordance since the form or shape of the machine must not directly relate to the human requirement. A calculator wants to be used because most humans find it way too time-consuming and exhausting to calculate in the head or to use pen and paper. The GPS in the car wants to be used because people are just about to lose the ability to navigate roads. The heater wants to be used when it is cold. Machines directly address human needs and frequently more than one. A car allows for transport from point a to b but may also represent social status. In addition, a car wants to be used or it breaks down (battery, etc). The artificial superintelligence

wants to be used all the time because it simplifies and organises human lives, reduces efforts and satisfies human needs. This is called universal solicitation.

By way of example, a military commander in a combat situation may step away from the use of autonomous, weaponised drones guided by artificial intelligence, nevertheless, the knowledge of the technology, its availability and impact will influence future decisions. A person may not actually buy or use a sex robot, however, knowledge of the existence of the devices, imagery associated with the machines and their availability will have an impact on relationships. If someone chooses to live a simple life in a remote area without any high tech, then this choice is the result of the existence of the technology, even if the recluse has no direct exposure to it. There is still the solicitation.

The vision here is a form of advanced artificial intelligence that is not just “human controlled” but beneficial in the sense that it explains itself and its operation to everybody. This includes the most vulnerable in a society, including children, the elderly and persons with an intellectual disability. These explanations must be in a form suitable and acceptable to users. Most probably, these explanations will take the form of videos and demonstrations. Even advanced forms of AI needs to match human comprehension.

Given significant advancements in brain-computer interfaces and the prospect of a direct link between neural and electronic information processing, a number of principles are important from a psychological point of view. The human motivational system must remain operational, in particular the reward system. Motivation breaks down if rewards are delivered to the human brain directly with no action or effort required to gain these rewards. This imposes restrictions on brain-machine interfaces. When the enhancement of cognitive abilities becomes possible by technical means, and at a stage when rewards are delivered directly into the human brain, the risk is that we “do not have to do anything anymore to get something”. This is certainly a challenge to human existence.

For those who are not convinced by the concepts of a human controllable AI, there must be the right to withdraw and to live a simpler life. Currently, the Amish people in the USA and Canada refuse the use of advanced technology and live compatibly with the societies surrounding these communities. There needs to be an opportunity to live away from artificial intelligence while still having the option to benefit from some aspects. Many people may seek a simpler life but who does not want to take advantage of progress in healthcare and the delivery of medical services?

The rapid development of artificial intelligence requires a very complex set of decisions for many people; for those who develop and use the technology but also for those who prefer to remain untouched. These decisions are deeply personal in nature. This book aims to provide some of the information required for this decision-making process.

I would like to thank my wife Susan Kay Wright for her love and support as well as significant feedback on early drafts of this book. Luke Diederich, Leonie Holthaus and Peter Trawny provided valuable comments and suggestions.

Queensland, Australia

Joachim Diederich

Contents

- 1 Universal Solicitation** 1
 - 1.1 Introduction 1
 - 1.2 Artificial Superintelligence 2
 - 1.2.1 The Singularity 3
 - 1.3 Regulating Societies by Use of Artificial Intelligence 5
 - 1.4 Robotics and AI 7
 - 1.5 Speed of Communication: Human vs AI 8
 - 1.6 Should We Have This Discussion Now? 9
 - 1.7 Will a Future AI Superintelligence Be Hostile? 9
 - 1.8 Can an Advanced AI Be Beneficial? 10
 - 1.9 Motivation and Background 11
 - 1.10 Overview 13
 - References 13
- 2 Digital Clones** 15
 - 2.1 Introduction 15
 - 2.2 Design Principles 16
 - 2.3 Psychotherapy as Heuristic Search 19
 - 2.3.1 Introduction to Heuristic Search 20
 - 2.3.2 Motivation for the Use of Heuristic Search 21
 - 2.4 Live Training of Digital Clones 22
 - 2.5 Digital Clone Trees 23
 - 2.6 Computers Are Social Actors 24
 - 2.7 Adjusting Language in Psychotherapy 25
 - 2.8 The Solicitation of AI Therapy 26
 - 2.9 Loneliness 28
 - 2.10 The “Caring Professions” and the Future of AI 30
 - References 30
- 3 Explanation** 33
 - 3.1 Introduction 33
 - 3.2 The Logic of Explanation 34
 - 3.3 Explanation and Cognition 35

3.4	Explanation and Learning	36
3.4.1	Explanation-Based Generalization	37
3.5	Types of Explanation	39
3.5.1	How and Why Explanations	40
3.5.2	Generating or Identifying the Best Explanation	40
3.5.3	Explanation and the “Theory of Mind”	41
3.6	Explanation and Black Box Machine Learning	42
3.6.1	Rule Extraction from Black-Box Machine Learning Systems	43
3.6.2	Rule-Extraction for Whom?	47
3.7	Visualisation and What-If Explanation	48
3.8	New Forms of Explanation	49
3.9	Explanation for Children	50
3.9.1	Satisfying Explanations for Children	50
3.10	Irrational Explanations	51
	References	52
4	Transhumanism	55
4.1	Introduction	55
4.2	Cognitive and Perceptual Enhancement	58
4.3	The Convergence of Abilities	60
4.3.1	Enhancements, Personality and the Notion of “Self”	61
4.4	The Concept of Self	61
4.4.1	Self Memory Systems	63
4.4.2	The Self and Autopoiesis	64
4.4.3	Enhancement and Social Cohesion	66
4.5	Brain-Machine Interfaces	66
4.5.1	BMI, Telepathy and Psychosis	68
4.6	Human Motivation	69
4.7	The Extension of Life	70
	References	71
5	Neo-Luddism	73
5.1	Introduction	73
5.2	Neo-Luddites	75
5.3	“Only a God Can Save Us”: Heidegger’s Spiegel Interview in 1966	77
5.3.1	Artificial Intelligence Is Not (Just) a Tool	81
5.4	A Thought Experiment	82
5.4.1	Mindfulness in Psychology	83
5.4.2	Husserl’s Phenomenological Reduction	84
5.4.3	Heideggerian AI	85
5.5	Resistance to Artificial Intelligence	85
5.5.1	Lethal Autonomous Weapons	85
5.5.2	Monitoring Online Conversations	87
5.5.3	AI and Privacy	88

5.5.4	Resistance to Medical AI	88
5.5.5	Determination of Ethnicity, Personality and Attractiveness	90
5.6	Societal Response to Advanced Technology	91
	References	92
6	Safety and Military Artificial Intelligence	95
6.1	Introduction	95
6.2	Attitudes Towards Military AI in the USA	96
6.3	Are Machines Responsible Agents?	98
6.3.1	Are Intelligent Killing Machines Responsible Agents?	99
6.3.2	Who Contributes to Killing Humans by Use of AI?	100
6.3.3	Nazi Atrocities, Liability and Personal Guilt	100
6.3.4	Climate Change and Personal Guilt	101
6.3.5	Non-humanoid Drones and Anxiety	102
6.4	Emotion and Military AI	102
6.5	Explanation and Military AI	103
6.5.1	Explanation and Misinformation in Military Applications	104
6.6	The Rules of Armed Conflict	106
6.6.1	The Rules of Armed Conflict and AI	107
6.6.2	An Extended Set of Rules for Armed AI Conflicts	107
6.7	Human Control of Advanced AI	109
6.7.1	Human AI Partnerships	109
6.7.2	Intrusion	109
6.8	Artificial General Intelligence and Military AI	112
	References	113
7	Risks to Artificial Intelligence	115
7.1	Introduction	115
7.2	The Uncanny Valley of Humans and Machines	118
7.2.1	Challenging the Uncanny Valley Concept	118
7.3	The Uncanny Valley and Autism Spectrum Disorder	119
7.3.1	Life in the Uncanny Valley	120
7.3.2	Theory of Mind, Autism and Artificial Superintelligence	121
7.4	Abuse of Chatbots	122
7.4.1	Porn Filters in Chatbots	123
7.5	Robots for Autism	124
	References	125
	Appendices	127