Volume 41 Number 4 December 2017 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Special Issue:

Superintelligence

Guest Editors:

Ryan Carey
Matthijs Maas
Nell Watson
Roman Yampolskiy

Editorial Boards

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Editor-in-Chief's Introduction to the Special Issue on "Superintelligence", AI and an Overview of IJCAI 2017

This editorial consists of two parts: first, an introduction to the Superintelligence special issue and, second, the traditional AI and IJCAI overview, which this year has been a little delayed.

1 Superintelligence special issue

Being Editor-In-Chief means many tedious hours of rapid proof-reading for all the papers, several times, as well as choosing interesting special issues and writing editorials. One of the great joys of the editorial work is to see an excellent special issue delivered, like this one on superintelligence. Let me shed some light on the editorial part of the special issue.

First of all, with help from colleagues at the Future of Life Institute we came across Roman Yampolskiy, born in Latvia and a graduate of the University of Buffalo. His book "Artificial Superintelligence: A Futuristic Approach" (Figure 1) represents a more technically oriented viewpoint than Bostrom's philosophical "Superintelligence: Paths, Dangers, Strategies". He is an associate at the Global Catastrophic Risk Institute, a think tank that analyses global risks to the survival of human civilization. Whatever the case, Roman gave a major boost to the superintelligence special issue.

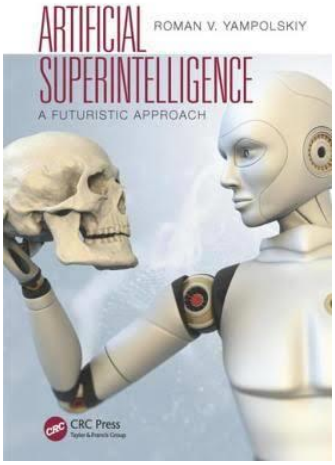


Figure 1: Yampolskiy's book on Artificial Superintelligence.

Three more special editors joined in the following weeks: Nell Watson, Matthijs Maas and Ryan Carrey.

Nell Watson is an engineer and futurist thinker, a kind of Northern Irish Ray Kurzweil, the latest probably the best-known singularity enthusiast and forecaster. She also advises The Lifeboat Foundation, helping to pinpoint the trajectories of human progress. She works at the Singularity University and the United Nations University.

Matthijs Maas is a PhD Fellow in Law and Policy on Global Catastrophic and Existential Threats at the University of Copenhagen's Centre for International Law, Conflict and Crisis (CILCC). He holds an M.Sc. in International Relations from the University of Edinburgh. His research interests include the safe governance of artificial intelligence and the effects of emerging technologies on strategic stability, amongst others. He is also a Junior Associate of the Global Catastrophic Risk Institute.

Ryan Carey is a research intern in AI safety at Ought Inc and a research affiliate at the Centre for Study of Existential Risk. He edited the Effective Altruism Handbook, a compilation of essays about how to do more good with limited resources. He also founded the Effective Altruism Forum and cofounded Effective Altruism Melbourne.

While the quality of any journal issue comes down to the authors of the papers, the editors of the special issue deserve particular attention as well.

Last but not least, let me thank Tine Kolenik, a student that helped me with this special issue, and Drago Torkar, technical editor of Informatica, for special efforts with the editorial system.

2 AI and IJCAI 2017

The progress of artificial intelligence (AI) is certainly fast and furious from the technical point of view. Each year there are scores of new achievements in academia, gaming, industry, and real life, having implications for the way we live and work. For example, autonomous vehicles are improving constantly, and are being introduced into more and more countries. Recently, IJCAI presented its annual general overview of the AI SOTA and progress.

Figure 2: Increase in the number of IJCAI papers in recent years.

The 26th International Joint Conference on Artificial Intelligence was held in Melbourne, Australia in August 2017 [6]. Melbourne has been judged the world's most liveable city for the seventh year running and indeed it is safe, clean, uncrowded, full of green nature and architectural wonders. It is a prosperous city that hosted a prosperous scientific event!

The growth in AI is indicated by the number of papers submitted to the IJCAI conference (Figure 2). In

2016 in New York there were 2294 papers, while in 2017 in Melbourne, 2540 papers were reviewed. The growth has been steady since 2009.

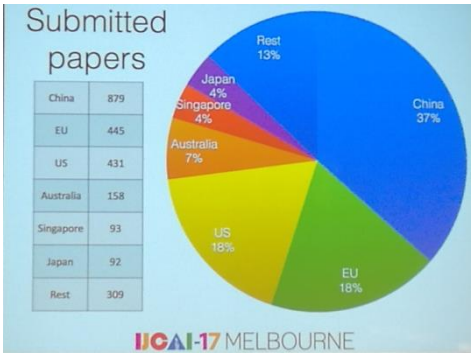


Figure 3: Papers per country at IJCAI 2017.

A study of the papers submitted per country (Figure 3) at IJCAI 2017 indicates that the majority was from China (37%), with the EU second (18%) and the US third (18%).

Although there was a lack of Eastern-European papers and papers from Russia, on September 1 Vladimir Putin, speaking with students, warned that whoever cracks AI will 'rule the world' [9]. Will that be China, as it already submits the most AI papers? Among others, at least 600 top Chinese ministers and military officials use quantum-encrypted links for all confidential communication. But the current leader is still USA with most of the awards given to the USA researchers at IJCAI 2017.

While the number of AI papers from China might come as a surprise, their industry achievements are astonishing as well. One might not be as familiar with the Chinese solutions as with Google or Amazon AI systems, but the Chinese systems are close to the top. For example, in 2017 China's Alibaba Group Holding Ltd introduced a cut-price voice-assistant speaker, similar to Amazon.com Inc's "Echo". It is called "Tmall Genie" and costs $73, significantly less than the western counterparts by Amazon and Alphabet Inc's Google, which cost around $150. Similarly, Baidu, China's top search engine, recently launched a device based on its own Siri-like "Duer OS" system. Alibaba and China's top tech firms have ambitions to become world leaders in AI.

In 2017, two games stood out as another example of AI beating the best human counterparts: unlimited Texas hold'em poker (10 on 160 possibilities) and Dota 2. Both games were slightly limited – in poker, there are only two players instead of more, and Dota 2 was also reduced to only two players instead of 10. Nevertheless, both games are the most-played human games with award funds going into the tens of millions. Both games are quite different from formal games like chess or Go. For example, poker includes human-bluffing interactions and hidden cards. Dota 2 is another surprise since it resembles fighting in the real world, although everything is more of a fantasy story. The key components were strategic plans with global and local decision making, and adapting to the adversary. From Wikipedia: "Dota 2 is originally played in matches between two teams of five players, with each team occupying and defending their own separate base on the map. Each of the ten players independently controls a powerful character, known as a "hero", who all have unique abilities and differing styles of play. During a match, the player collects experience points and items for their heroes in order to successfully fight the opposing team's heroes, who are doing the same. A team wins by being the first to destroy a large structure located in the opposing team's base, called the "Ancient", which is guarded by defensive towers."

Regarding the methods, reinforcement learning and deep neural networks were the most commonly applied; however, the AI field at IJCAI 2017 was presented for more than 10 major areas.

Various types of deep neural networks (DNNs) continue their excellence in visual recognition tasks and in real-life diagnostics, such as diagnosing which tissue contains malignant cancer cells. When fed with huge numbers of examples and with fine-tuned parameters, DNNs regularly beat the best human experts in increasing numbers of artificial and real-life tasks, like diagnosing tissue in several diseases. There are other everyday tasks, e.g., the recognition of faces from a picture, where DNNs recognized hundreds of faces in seconds, a result no human can match. Figure 4 demonstrates the progress of DNNs in visual tasks: around 2015 the visual recognition in specific domains was comparable to humans; now, it has surpassed humans quite significantly – again, in particular visual tests. BTW, DNNs are currently breaking the CAPTCHA test – the simplest way so far to differentiate between SW agents and humans.

The effects of only visual superiority are astonishing on their own, but several services emerge from visual analyses. For example, eye analyses make it possible to detect certain diseases like cancer or Alzheimer's [3]. Furthermore, DNN studies of facial properties can reveal sexual orientation, IQ, and political orientation. When shown five photos of a man, a recent system was able to correctly select the man's sexuality 91 per cent of the time, while humans were able to perform the same task with less than 70% accuracy [7]. This Stanford University study alone confirmed that homosexuality is very probably of genetic origin. The consequences of a single study can be profound. Will job applications also be assessed by a DNN study of facial properties? Will dictatorships prosecuting homosexuality punish their citizens on the basis of their faces?

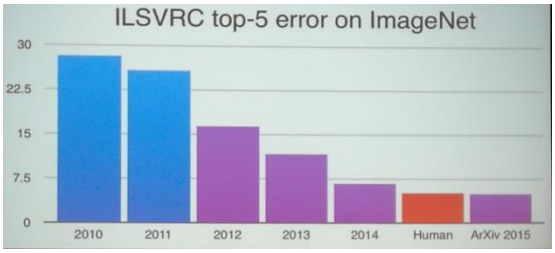


Figure 4: Error of DNNs on ImageNet over time.

There were several demonstrations and competitions at IJCAI 2017, including the traditional Angry Birds' competition. Most attractive, however, were soccer competitions with off-line Nao robots that were not trained or advised as a team, but performed on their own in a group decided on the spot. Unfortunately, the local computing powers are at the level of a mobile phone, which means they are insufficient for good play. The robots were often wandering around, searching for the ball. Still, they demonstrated some quite cunning skills, e.g., accurately kicking the ball into the goal using a specific angle relative to the foot.

Next year will be of particular interest. ICML with 3500 attendees, IJCAI+ECAI with 2500, AAMAS with 700, ICCBR with 250 and SOCS with 50 attendees will be hosted in Stockholm in a 2-week event in July 2018. Optimistic jokes are emerging that the critical mass of 6–7000 attendees will be sufficient to ignite general intelligence or even a path to superintelligence [2, 8, 10].

As part of the tradition, the IJCAI overview avoids mentioning people's names; however, there is one person that deserves special attention – Toby Walsh, Local Arrangements Committee Co-Chair, a key organiser of the letter "Killer robots: World's top AI and robotics companies urge United Nations to ban lethal autonomous weapons" and organizer of Melbourne's public Festival of Artificial Intelligence. Congratulations!

There are two additional matters worth mentioning: the ban on autonomous weapons and the Asilomar principles.

2.1 Ban on autonomous weapons

There are two major reasons for the proposed ban:

- Fully autonomous weapons will likely make war inhumane, whereas humans – if war cannot be avoided – need some rule of engagement to preserve some level of humanity and prevent human suffering being too extreme.
- This is one of the preconditions on the road to prevent superintelligence from going viral and malignant [2, 8, 10].

There is good reason for celebrating the first successes of the pro-ban efforts – the movement is spreading through social media since it started years ago by scientists like Toby Walsh or Stuart Russel and is currently coordinated by Mary Wareham.

Slovenia is involved in the ban at the European and national levels, where four societies (SLAIS for artificial intelligence, DKZ for cognitive science, Informatica for informatics, ACM Slovenia for computer science) drew up a letter and sent it to the UN and Slovenian government, and recently the Slovenian AI society SLAIS wrote a letter to the European national communities to join activities in this direction. Our initiative was also debated at the European AI society EurAI meeting at IJCAI 2017.

Second, Elon Musk and the CEOs of 155 robotic companies signed a letter in which they say "Once developed, lethal autonomous weapons will permit armed conflict to be fought at a scale greater than ever, and at timescales faster than humans can comprehend. These can be weapons of terror, weapons that despots and terrorists use against innocent populations, and weapons hacked to behave in undesirable ways."

"We do not have long to act. Once this Pandora's Box is opened, it will be hard to close."

On the other hand, the world's superpowers are rapidly not only developing, but also applying autonomous weapons, from drones to tanks or submarines. Some even argue that it is already too late to stop these autonomous weapons.

Another example: the EU parliament accepted new legislation giving artificial systems some of the rights of living beings. This is exactly one of the rules of thumb that should not be done to avoid potentially negative AI progress. So why did EU politicians accept such a law? It is not dangerous yet, but clearly worrisome.

2.2 The 23 Asilomar principles

The Future of Life Institute's [4] second conference on the future of AI was organized in January 2017. The purpose of this section is to introduce, in the rather original way, the 23 Asilomar AI principles [1] defined at the BAI 2017 conference.

The opinion of the BAI 2017 attendees and the world-wide AI community is widely held: "a major change is coming, over unknown timescales but across every segment of society, and the people playing a part in that transition have a huge responsibility and opportunity to shape it for the best." Therefore, a list of Asilomar principles was designed to provide directions for future AI research.

The first task of the organizers was to compile a list of scores of opinions about what society should do to best manage AI in the coming decades. From this list, the organizers distilled as much as they could into a core set of principles that expressed some level of consensus. The coordinating effort dominated the event, resulting in a significantly revised version for use at the meeting. There, small breakout groups discussed subsets of the principles, giving detailed refinements and commentaries on them. This process generated improved versions of the principles. Finally, they surveyed the full set of attendees to determine the level of support for each version of each principle.

After this time-consuming and meticulous process, a high level of consensus emerged around many of the statements during the final survey. The final list retained principles that at least 90% of the attendees agreed on. The 23 principles were grouped into research strategies, data rights and future issues including potential superintelligence, signed by those wishing to associate their name with the list. The principles with additional interviews can be obtained from the web pages of the event at the Future of Life Institute [4]. The principles will hopefully provide some guidelines as to how the power of AI can be used to improve everyone's lives in future years.

AI has already provided useful tools that are employed every day by people all around the world. Its

continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead.

3 Conclusion

AI's progress is both fascinating and accelerating. An increasing awareness of AI-related changes in human society is being recognised by the scientific, academic and general public. Dozens of major reports have emerged from academia (e.g., the Stanford 100-year report), government (e.g., two major reports from the White House), industry (e.g., materials from the Partnership on AI), and the non-profit sector (e.g., a major IEEE report). The special issue on superintelligence will hopefully spur discussion and awareness among the public, media and government, helping them to understand that the times are changing rapidly, and that new approaches and methods are needed for humans to successfully cope with the future.

On the other hand, AI stubbornly lacks general intelligence and other human properties like consciousness. There are specific claims that current computers do not provide the kind of computing that can emulate the best human intellectual properties [5]. The Turing test remains as a mission impossible for even the most advanced systems. It may be that we do not need to worry so much about overall global superintelligence bypassing humans in every category, but to focus on the technical progress of AI and its applications.

Scientific understandings about AI, its influence on everyday life, and the future of human civilization are stacking up. Scientists are able to provide some guidelines about which direction we humans should develop AI to avoid the dangers of the negative effects of the rising power of AI. While AI often frightens the general public, I, and several AI researchers, find its rapid progress a necessity to prevent the decline or even the self-destruction of human civilization. These potential dangers are real, not fictitious, primarily because of the simple fact that any major power can be easily misused to cause harm to humans, and second, there are some strong indications that civilizations tend to destroy themselves (the Fermi paradox). By raising awareness, we increase the chances to reap the positive aspects of an amazing future AI and avoid the negative ones.

References

[1] Asilomar principles. 2017, (https://futureoflife.org/2017/01/17/principled-ai-discussion-asilomar/).

[2] Bostrom, N. 2014. Superintelligence – Paths, Dangers, Strategies. Oxford University Press, Oxford, UK.

[3] Eye Scans to Detect Cancer and Alzheimer's Disease, https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/eye-scans-to-detect-cancer-and-alzheimers-disease

[4] Future of life institute, https://futureoflife.org/

[5] Gams, M. 2001. Weak intelligence: through the principle and paradox of multiple knowledge. Nova Science.

[6] IJCAI conference, 2017, https://ijcai-17.org

[7] Kosinski, M., Wang. Y. 2017. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. https://osf.io/zn79k/

[8] Kurzweil, R. 2006. The Singularity Is Near: When Humans Transcend Biology, Sep 26, Penguin Books.

[9] Mail online, Science and technology, Vladimir Putin warns whoever cracks artificial intelligence will 'rule the world', http://www.dailymail.co.uk/sciencetech/article-4844322/Putin-Leader-artificial-intelligence-rule-world.html

[10] Yampolskiy, R.V. 2016. Artificial Superintelligence. CRC Press.

Matjaž Gams

Guest Editors' Introduction to the Special Issue on "Superintelligence"

The concept of superintelligence was developed only a little after the birth of the field of artificial intelligence, and it has been a source of persistent intrigue ever since. Alan Turing himself toyed with the idea of human-level intelligence: "If a machine can think, it might think more intelligently than we do, and then where should we be?"[1] In 1965, I.J. Good, a former colleague of Turing's, considered what would happen if a machine could effectively redesign itself.[2] This, he argued, could lead to what we would now call a superintelligence: a system that "greatly exceeds the cognitive performance of humans in virtually all domains of interest".[3]

In some ways, our understanding of how to deal with these problems has advanced little since then. Although the field of artificial intelligence has advanced substantially, it is quite unclear by what pathway superintelligence may be reached. Given that technological forecasting is covered in such a haze, we cannot say that superintelligent AI will come soon, but neither can we be assured that it will be far away. It would be similarly complacent to claim to know with confidence whether such a system will be beneficial or harmful by default. It is troubling that we still find ourselves so uncertain about these 'crucial considerations'[4]: the emergence of 'human' intelligence proved a watershed in the history of the earth (certainly in our history), and the prospective development of superintelligence is unlikely to be any smaller in its impact and ramifications. Now may be (and is on expectation), a critical time to think more, so that we can see a sharper outline of this situation, and formulate plan for managing it.

Over the last decade, a range of academics have finally begun to respond to this challenge in an organized way. Many core philosophical issues have been charted, and technical AI safety research is now an emerging field;[5] there is also a young but ambitious research agenda exploring the geopolitical impacts and governance challenges surrounding the eventual deployment of superintelligent systems.[6] Some of this research can find a natural home in the fields of computer science, political science, or philosophy. Much, however, cannot. The considerations in evaluating and planning for superintelligence often cut across the practical and the philosophical, the technical and the non-technical, riding across several academic disciplines such that the most important work will often have no natural home in any of them. So the purpose of this Special Issue is to collect some of these essays, that use a full range of tools to evaluate and plan for superintelligence.

We hope that this will generate insights and debates that can help us get a better handle on this important topic--to enable us to undertake the conceptual, technological and societal innovations that will make superintelligence beneficial for the world.

The contributions to this issue can be coarsely separated into two baskets. Four of the contributions primarily focus on improving our understanding of the strategic landscape: they characterise the development of superintelligence, and its potential consequences. The remaining three chart a path through this landscape: they argue for specific kinds of research in order to make beneficial outcomes more likely.

In 'Superintelligence as a cause or cure for risks of astronomical suffering', **Kaj Sotala & Lukas Gloor** outline a new category of "suffering risks" ('s-risks'), in which astronomical suffering occurs on an astronomical scale. They propose that such risks may be of comparable severity and probability as extinction risks, and survey some of the ways that superintelligent AI might either bring about or relieve these kinds of risks, and some of the ways that further theoretical work could affect these.

In 'Artificial Intelligence in Life Extension: from Deep Learning to Superintelligence', **Michael Batin, Alexey Turchin, Sergey Markov, Alisa Zhila, David Denkenberger** consider how steadily advancing AI could be used to extend the human lifespan. They offer an extensive survey of presently ongoing and potential future AI applications to anti-aging research, at three stages of development--narrow AI, AGI, and superintelligence, finding that medical-focused superintelligence might help humans to achieve 'longevity escape velocity'.

In 'Modeling and Interpreting Expert Disagreement About Artificial Superintelligence', **Seth Baum, Anthony Barrett, and Roman Yampolskiy** consider how we are to deal with persistent, pervasive expert disagreement about the risks posed by superintelligence. They describe a 'ASI-PATH' fault-tree model, and use it

[1] Turing, Alan. "Can digital computers think?(1951)." B. Jack Copeland (2004): 476.

[2] Good, I. J. "Speculations Concerning the First Ultraintelligent Machine*." Edited by Franz L. Alt and Moris Rubinoff. Advances in Computers 6 (1965): 31–88.

[3] Bostrom, 2014: 25.

[4] Bostrom, Nick. 2014. *Crucial considerations and wise philanthropy.*

[5] See agendas for this work at [Taylor, Jessica, et al. "Alignment for advanced machine learning systems." Machine Intelligence Research Institute (2016).] and [Amodei, Dario, et al. "Concrete problems in AI safety." arXiv preprint arXiv:1606.06565 (2016).], and collections that include some of this work (as well as some other kinds) at https://futureoflife.org/ai-safety-research/ and http://effective-altruism.com/ea/1iu/2018_ai_safety_literature_review_and_charity/.

[6] Recent work in this area is compiled at http://www.allandafoe.com/aireadings

to chart points of disagreement between Nick Bostrom and Ben Goertzel, over the viability of catastrophic risks from superintelligence. They show how this model can assist with weighing the importance of different considerations, and can help with prioritization of superintelligence risk management strategies.

David Jilk, in 'Conceptual-Linguistic Superintelligence' reviews the ingredients and dynamics of an 'intelligence explosion', arguing that any AI system capable of sustaining such an intelligence explosion must have a 'conceptual-linguistic' faculty with functional similarity to that found in humans.

The papers that proposed future research were quite complementary to one another. Each of the three proposed a kind of research that could draw on different kinds of expertise to the others.

Gopal Sarma & Nick Hay, in 'Mammalian Value Systems', seek to bring fresh insights from other academic disciplines to bear on the problem of aligning AI goals with human values. They argue that what we call human values can be decomposed into (1) mammalian values, (2) human cognition, (3) human social and cultural evolution. They further argue that having more detailed prior information on the structures of human values may enable AI agents to infer these values from fewer examples, and advocate, on this basis, for greater research on mammalian values.

In their second submission, 'Robust computer Algebra, Theorem Proving and Oracle AI**', Sarma & Hay** provide another concrete avenue for AI safety research. They identify 'computer algebra systems' (CAS) as primitive examples of domain-specific oracles.; By charting efforts to integrate such computer algebra systems with theorem provers, they lay out a concrete set of encountered problems and considerations relevant to the 'provable safety' of eventual superintelligent 'Oracle AI'.

In 'The Technological Landscape Affecting Artificial General Intelligence and the Importance of Nanoscale Neural Probes', **Daniel Eth** argues that the development of nanoscale neural probes could substantially increase the likelihood that whole brain emulations are the first kind of AGI developed (as opposed to 'de novo' AI and neuromorphic AI). He argues as a result that it is desirable for research effort to be dedicated to accelerating their development.

Although the study of superintelligence has resurged in the last decade, it is still at a relatively early stage of maturity. It is one of the most exciting--and plausibly one of the most important--research areas of our time. As guest editors, we hope to have collected some work that has shone a small light on some of the problem of what to do about superintelligence. We are grateful to all authors for their contributions to this issue , and for their broader work exploring this critical topic. We also give special thanks to Prof. Matjaz Gams, Editor-in-chief of Informatica, for his support in composing this special issue.

Ryan Carey

Matthijs Maas

Nell Watson

Roman Yampolskiy

Superintelligence as a Cause or Cure for Risks of Astronomical Suffering

Kaj Sotala and Lukas Gloor
Foundational Research Institute, Berlin, Germany
E-mail: kaj.sotala@foundational-research.org, lukas.gloor@foundational-research.org
foundational-research.org

Keywords: existential risk, suffering risk, superintelligence, AI alignment problem, suffering-focused ethics, population ethics

Received: August 31, 2017

Discussions about the possible consequences of creating superintelligence have included the possibility of existential risk, often understood mainly as the risk of human extinction. We argue that suffering risks (s-risks), where an adverse outcome would bring about severe suffering on an astronomical scale, are risks of a comparable severity and probability as risks of extinction. Preventing them is the common interest of many different value systems. Furthermore, we argue that in the same way as superintelligent AI both contributes to existential risk but can also help prevent it, superintelligent AI can be both the cause of suffering risks and a way to prevent them from being realized. Some types of work aimed at making superintelligent AI safe will also help prevent suffering risks, and there may also be a class of safeguards for AI that helps specifically against s-risks.

Povzetek: Prispevek analizira prednosti in nevarnosti superinteligence.

1 Introduction

Work discussing the possible consequences of creating superintelligent AI (Yudkowsky 2008, Bostrom 2014, Sotala & Yampolskiy 2015) has discussed superintelligence as a possible *existential risk*: a risk "where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential" (Bostrom 2002, 2013).

The previous work has mostly[1] considered the worst-case outcome to be the possibility of human extinction by an AI that is indifferent to humanity's survival and values. However, it is often thought that for an individual, there exist "fates worse than death"; analogously, for civilizations there may exist fates worse than extinction, such as survival in conditions in which most people will experience enormous suffering for most of their lives.

Even if such extreme outcomes would be avoided, the known universe may eventually be populated by vast amounts of minds: published estimates include the possibility of 10^{25} minds supported by a single star (Bostrom 2003a), with humanity having the potential to eventually colonize tens of millions of galaxies (Armstrong & Sandberg 2013). While this could enable an enormous number of meaningful lives to be lived, if even a small fraction of these lives were to exist in hellish circumstances, the amount of suffering would be vastly greater than that produced by all the atrocities, abuses, and natural causes in Earth's history so far.

We term the possibility of such outcomes a *suffering risk:*

> Suffering risk (s-risk): One where an adverse outcome would bring about severe suffering on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far.

In order for potential risks - including s-risks - to merit work on them, three conditions must be met. First, the outcome of the risk must be sufficiently severe to merit attention. Second, the risk must have some reasonable probability of being realized. Third, there must be some way for risk-avoidance work to reduce either the probability or severity of an adverse outcome.

In this paper, we will argue that suffering risks meet all three criteria, and that s-risk avoidance work is thus of a comparable magnitude in importance as work on risks from extinction. Section 2 seeks to establish the severity of s-risks. There, we will argue that there are classes of suffering-related adverse outcomes that many value systems would consider to be equally or even more severe than extinction. Additionally, we will define a class of less severe suffering outcomes which many value systems would consider important to avoid, albeit not as important as avoiding extinction. Section 3 looks at suffering risks from the view of several different value systems, and discusses how much they would prioritize avoiding different suffering outcomes. Next, we will argue that there is a reasonable probability for a number of different suffering risks to be realized. Our discussion is organized according to the relationship that superintelligent AIs have to suffering risks: section 4 covers risks that may be prevented by a

[1] Bostrom (2014) is mainly focused on the risk of extinction, but does also devote some discussion to alternative negative outcomes such as "mindcrime". We discuss mindcrime in section 5.

superintelligence, and section 5 covers risks that may be realized by one[2]. Section 6 discusses how it might be possible to work on suffering risks.

2 Suffering risks as risks of extreme severity

As already noted, the main focus in discussion of risks from superintelligent AI has been either literal extinction, with the AI killing humans as a side-effect of pursuing some other goal (Yudkowsky 2008), or a *value extinction*. In value extinction, some form of humanity may survive, but the future is controlled by an AI operating according to values which all current-day humans would consider worthless (Yudkowsky 2011). In either scenario, it is thought that the resulting future would have no value.

In this section, we will argue that besides futures that have *no value*, according to many different value systems it is possible to have futures with *negative value*. These would count as the worst category of existential risks. In addition, there are adverse outcomes of a lesser severity, which depending on one's value systems may not necessarily count as worse than extinction. Regardless, making these outcomes less likely is a high priority and a common interest of many different value systems.

Bostrom (2002) frames his definition of extinction risks with a discussion which characterizes a single person's death as being a risk of terminal intensity and personal scope, with existential risks being risks of terminal intensity and *global* scope - one person's death versus the death of all humans. However, it is commonly thought that there are "fates worse than death": at one extreme, being tortured for an extended time (with no chance of rescue), and then killed.

As less extreme examples, various negative health conditions are often considered worse than death (Rubin, Buehler & Halpern 2016; Sayah et al. 2015; Ditto et al., 1996): for example, among hospitalized patients with severe illness, a majority of respondents considered bowel and bladder incontinence, relying on a feeding tube to live, and being unable to get up from bed, to be conditions that were worse than death (Rubin, Buehler & Halpern 2016). While these are prospective evaluations rather than what people have actually experienced, several countries have laws allowing for voluntary euthanasia, which people with various adverse conditions have chosen rather than go on living. This may considered an empirical confirmation of some states of life being worse than death, at least as judged by the people who choose to die.

The notion of fates worse than death suggests the existence of a "hellish" severity that is one step worse than "terminal", and which might affect civilizations as well as individuals. Bostrom (2013) seems to acknowledge this by including "hellish" as a possible severity in the corresponding chart, but does not place any concrete outcomes under the hellish severity, implying that risks of extinction are still the worst outcomes. Yet there seem to be plausible paths to civilization-wide hell outcomes as well (Figure 1), which we will discuss in sections 4 and 5.

Global	Thinning of the ozone layer	**Extinction risks**	**Global hellscape**
Personal	Car is stolen	Death	Extended torture followed by death
	Endurable	**Terminal**	**Hellish**

Figure 1: The worst suffering risks are ones that affect everyone and subject people to hellish conditions.

In order to qualify as equally bad or worse than extinction, suffering risks do not necessarily need to affect every single member of humanity. For example, consider a simplified ethical calculus where someone may have a predominantly happy life (+1), never exist (0), or have a predominantly unhappy life (-1). As long as the people having predominantly unhappy lives outnumber the people having predominantly happy lives, under this calculus such an outcome would be considered worse than nobody existing in the first place. We will call this scenario a **net suffering outcome**[3].

This outcome might be considered justifiable if we assumed that, given enough time, the people living happy lives will eventually outnumber the people living unhappy lives. Most value systems would then still consider a net suffering outcome worth avoiding, but they might consider it an acceptable cost for an even larger amount of future happy lives.

On the other hand it is also possible that the world could become locked into conditions in which the balance would remain negative even when considering all the lives that will ever live: things would never get better. We will call this a **pan-generational net suffering outcome.**

In addition to net and pan-generational net suffering outcomes, we will consider a third category. In these outcomes, serious suffering may be limited to only a fraction of the population, but the overall population at some given time[4] is still large enough that even this small fraction accounts for many times more suffering than has

[2] Superintelligent AIs being in a special position where they might either enable or prevent suffering risks, is similar to the way in which they are in a special position to make risks of extinction both more or less likely (Yudkowsky 2008).

[3] "Net" should be considered equivalent to Bostrom's "global", but we have chosen a different name to avoid giving the impression that the outcome would necessarily be limited to only one planet.

[4] One could also consider the category of pan-generational astronomical suffering outcomes, but restricting ourselves into just three categories is sufficient for our current discussion.

existed in the history of the Earth. We will call these **astronomical suffering outcomes.**

Types of suffering outcomes	
Astronomical suffering outcome	At some point in time, a fraction of the population experiences hellish suffering, enough to overall constitute an astronomical amount that overwhelms all the suffering in Earth's history.
Net suffering outcome	At some point in time, there are more people experiencing lives filled predominantly with suffering than there are people experiencing lives filled predominantly with happiness.
Pan-generational net suffering outcome	When summed over all the people that will ever live, there are more people experiencing lives filled predominantly with suffering than there are people experiencing lives filled predominantly with happiness.

Figure 2: types of possible suffering outcomes. An outcome may count as one or several of the categories in this table.

Any value system which puts weight on preventing suffering implies at least some interest in preventing suffering risks. Additionally, as we will discuss below, even value systems which do not care about suffering *directly* may still have an interest in preventing suffering risks.

We expect these claims to be relatively uncontroversial. A more complicated question is that of *tradeoffs*: what should one do if some interventions increase the risk of extinction but make suffering risks less likely, or vice versa? As we will discuss below, if forced to choose between these two, different value systems will differ in which of the interventions they favor. In such a case, rather than to risk conflict between value systems, a better alternative would be to attempt to identify interventions which do not involve such a tradeoff. If there were interventions that reduced the risk of extinction without increasing the risk of astronomical suffering, or decreased the risk of astronomical suffering without increasing the risk of extinction, or decreased both, then it would be in everyone's interest to agree to jointly focus on these three classes of interventions.

3 Suffering risks from the perspective of different value systems

We will now take a brief look at different value systems and their stance on suffering risks, as well as their stance on the related tradeoffs.

Classical utilitarianism. All else being equal, classical utilitarians would prefer a universe in which there were many happy lives and no suffering. However, a noteworthy feature about classical utilitarianism (as well as some other aggregative theories) is that it considers very good and very bad scenarios to be symmetrical - that is, a scenario with 10^20 humans living happy lives may be considered equally good, as a scenario with 10^20 humans living miserable lives is considered bad.

Thus, people following classical utilitarianism or some other aggregative theory may find compelling the argument (Bostrom 2003a) that an uncolonized universe represents a massive waste of potential value, and be willing to risk - or even accept - astronomical numbers of suffering individuals if that was an unavoidable cost to creating even larger numbers of happiness. Thus, classical utilitarianism would consider astronomical and net suffering outcomes something to avoid but possibly acceptable, and pan-generational net suffering outcomes as something to avoid under all circumstances.

Other aggregative theories. Any moral theory which was not explicitly utilitarian, but still had an aggregative component that disvalued suffering, would consider suffering risks as something to avoid. Additionally, for moral theories that valued things other than just pleasure and suffering - such as preference satisfaction, some broader notion of "human flourishing", objective list theories - hellscape scenarios would likely also threaten the satisfaction of many of the things that these theories valued. For example, minds experiencing enormous suffering are probably not flourishing, are likely to have unsatisfied preferences, and probably do not have many of the things considered valuable in objective list theories.

Similarly to classical utilitarianism, many aggregative theories could be willing to risk or even accept astronomical and civilization-wide suffering outcomes as a necessary evil but wish to avoid pan-generational net suffering outcomes. At the same time, many aggregative theories might incorporate some suffering-focused intuition (discussed below) which caused them to put more weight on the avoidance of suffering than the creation of other valuable things. Depending on the circumstances, this might cause them to reject the kind of reasoning which suggested that suffering outcomes could be an acceptable cost.

Rights-based theories. Rights-based theories would consider suffering risks a bad thing *directly* to the extent that they held that people - or animals (Regan 1980) - had a right to be treated well avoid unnecessary suffering. They could also consider suffering risks *indirectly* bad, if the suffering was caused by conditions which violated some other right or severely constrained someone's capabilities (Nussbaum 1997, p. 287). For example, a right to meaningful autonomy could be violated if a mind was subjected to enormous suffering and had no meaningful option to escape it.

General suffering-focused intuitions. There are various moral views and principles which could fit many different value systems, all of which would imply that suffering risks were something important to avoid and which might cause one to weigh the avoidance of suffering more strongly than the creation of happiness:

1. *Prioritarianism.* Prioritarianism is the position that the worse off an individual is, the more morally valuable it is to make that individual better off (Parfit 1991). That is, if one person is living in hellish conditions and another is well-off, then making the former person

slightly better off is more valuable than improving the life of the well-off person by the same amount. A stance of "astronomical prioritarianism" that considers all minds across the universe, and prioritizes improving the worst ones sufficiently strongly, pushes in the direction of mainly improving the lives of those that would be worst off and thus avoiding suffering risks. If a suffering outcome does manifest itself, prioritarianism would prioritize bringing it to an end, over creating additional well-off lives or further helping those who are already well off. Prioritarianism may imply focusing particularly on risks from future technologies, as these may enable the creation of mind states that are worse than the current biopsychological limits.

Besides prioritarianism, the following three intuitions (Gloor & Mannino 2016) would also prioritize the avoidance of suffering risks[5]:

2. *Making people happy, not happy people*[6]. An intuition which is present in preference-based views such as antifrustrationism (Fehige 1998), antinatalism (Benatar 2008), as well as the "moral ledger" analogy (Singer 1993) and prior-existence utilitarianism (Singer 1993), is that it is more important to make existing people better off than it is to create new happy beings.[7] For example, given the choice between helping a million currently-existing people who are in pain and bringing ten million new people into existence, this view holds that it is more important to help the existing people, even if the ten million new people would end up living happy lives.

A part of this view is the notion that it is not intrinsically bad to never be created, whereas it is intrinsically bad to exist and be badly off, or to be killed against one's wishes once one does exist. If one accepts this position, then one could still want to avoid extinction - or at least the death of currently-living humans - but the promise of astronomical numbers of happy lives being created (Bostrom 2003a) would not be seen as particularly compelling, whereas the possible creation of astronomical numbers of lives experiencing suffering could be seen as a major thing to avoid.

3. *Torture-level suffering cannot be counterbalanced.* This intuition is present in the widespread notion that minor pains cannot be aggregated to become worse than an instant of torture (Rachels 1998), in threshold negative utilitarianism (Ord 2013), philosophical fictional works such as *The Ones Who Walk Away From Omelas* (LeGuin 1973), and it may contribute to the absolute prohibitions against torture in some deontological moralities. Pearce (1995) expresses a form of it when he writes, "No amount of happiness or fun enjoyed by some organisms can notionally justify the indescribable horrors of Auschwitz".

4. *Happiness as the absence of suffering*. A view which is present in Epicureanism as well as many non-Western traditions, such as Buddhism, is that of happiness as the absence of suffering. Under this view, when we are not experiencing states of pleasure, we begin to crave pleasure, and this craving constitutes suffering. Gloor (2017) writes:

Uncomfortable pressure in one's shoes, thirst, hunger, headaches, boredom, itches, non-effortless work, worries, longing for better times. When our brain is flooded with pleasure, we temporarily become unaware of all the negative ingredients of our stream of consciousness, and they thus cease to exist. Pleasure is the typical way in which our minds experience temporary freedom from suffering. This may contribute to the view that pleasure is the symmetrical counterpart to suffering, and that pleasure is in itself valuable and important to bring about. However, there are also (contingently rare) mental states devoid of anything bothersome that are not commonly described as (intensely) pleasurable, examples being flow states or states of meditative tranquility. Felt from the inside, tranquility is perfect in that it is untroubled by any aversive components, untroubled by any cravings for more pleasure. Likewise, a state of flow as it may be experienced during stimulating work, when listening to music or when playing video games, where tasks are being completed on auto-pilot with time flying and us having a low sense of self, also has this same quality of being experienced as completely problem-free. Such states - let us call them states of contentment - may not commonly be described as (intensely) pleasurable, but following philosophical traditions in both Buddhism and Epicureanism, these states, too, deserve to be considered states of happiness.

Under this view, happiness and pleasure are not intrinsically good, but rather *instrumentally* good in that pleasure takes our focus away from suffering and thus helps us avoid it. Creating additional happiness, then, has no intrinsic value if that creation does not help avoid suffering.

4 Suffering outcomes that could be prevented by a superintelligence

In the previous section, we argued that nearly all plausible value systems will want to avoid suffering risks and that for many value systems, suffering risks are some

[5] One might naturally also have various intuitions that point in the opposite direction, that is, of not prioritizing suffering risks. We will not survey these, as our intent in this section is merely to establish that many would consider suffering risks as important to avoid, without claiming that this would be the *only* plausible view to hold.

[6] The name of this intuition is a paraphrase of Narveson (1973), "We are in favor of making people happy, but neutral about making happy people."

[7] Moral views that attempt to incorporate this intuition by treating the creation of new people as morally neutral (e.g. Singer's "prior-existence" criterion) suffer from what Greaves (2017) calls a "remarkabl[e] difficult[y] to formulate any remotely acceptable axiology that captures this idea of 'neutrality'". The views by Benatar and Fehige avoid this problem, but they imply a more extreme position where adding new lives is neutral only in a best-case scenario where they contain no suffering or frustrated preferences.

of the worst possible outcomes and thus some of the most important to avoid. However, whether this also makes suffering risks the type of risk that is the most important to *focus on*, also depends on how probable suffering risks are. If they seem exceedingly unlikely, then there is little reason to care about them.

In this and the next section, we will discuss reasons for believing that there are various suffering outcomes that might realize themselves. We begin by considering outcomes which occur naturally but could be prevented by a superintelligence. In the next section, we will consider suffering outcomes which could be caused by a superintelligence.

A superintelligence could prevent almost any outcome if it established itself a singleton, "a world order in which there is a single decision-making agency at the highest level" (Bostrom 2005). Although a superintelligence is not the only way by which a singleton might be formed, alternative ways - such as a world government or convergent evolution leading everyone to adopt the same values and goals (Bostrom 2005) - do not seem particularly likely to happen soon. Once a superintelligence had established itself as a singleton, depending on its values it might choose to take actions that prevented suffering outcomes from arising.

4.1 Are suffering outcomes likely?

Bostrom (2003a) argues that given a technologically mature civilization capable of space colonization on a massive scale, this civilization "would likely also have the ability to establish at least the minimally favorable conditions required for future lives to be worth living", and that it could thus be assumed that all of these lives would be worth living. Moreover, we can reasonably assume that outcomes which are *optimized* for everything that is valuable are more likely than outcomes optimized for things that are disvaluable. While people want the future to be valuable both for altruistic and self-oriented reasons, no one intrinsically wants things to go badly.

However, Bostrom has himself later argued that technological advancement combined with evolutionary forces could "lead to the gradual elimination of all forms of being worth caring about" (Bostrom 2005), admitting the possibility that there could be technologically advanced civilizations with very little of anything that we would consider valuable. The technological potential to create a civilization that had positive value does not automatically translate to that potential being used, so a very advanced civilization could still be one of no value or even negative value.

Examples of technology's potential being unevenly applied can be found throughout history. Wealth remains unevenly distributed today, with an estimated 795 million people suffering from hunger even as one third of all produced food goes to waste (World Food Programme, 2017). Technological advancement has helped prevent many sources of suffering, but it has also created new ones, such as factory-farming practices under which large numbers of animals are maltreated in ways which maximize their production: in 2012, the amount of animals slaughtered for food was estimated at 68 billion worldwide (Food and Agriculture Organization of the United Nations 2012). Industrialization has also contributed to anthropogenic climate change, which may lead to considerable global destruction. Earlier in history, advances in seafaring enabled the transatlantic slave trade, with close to 12 million Africans being sent in ships to live in slavery (Manning 1992).

Technological advancement does not automatically lead to positive results (Häggström 2016). Persson & Savulescu (2012) argue that human tendencies such as "the bias towards the near future, our numbness to the suffering of great numbers, and our weak sense of responsibility for our omissions and collective contributions", which are a result of the environment humanity evolved in, are no longer sufficient for dealing with novel technological problems such as climate change and it becoming easier for small groups to cause widespread destruction. Supporting this case, Greene (2013) draws on research from moral psychology to argue that morality has evolved to enable mutual cooperation and collaboration within a select group ("us"), and to enable groups to fight off everyone else ("them"). Such an evolved morality is badly equipped to deal with collective action problems requiring global compromises, and also increases the risk of conflict and generally negative-sum dynamics as more different groups get in contact with each other.

As an opposing perspective, West (2017) argues that while people are often willing to engage in cruelty if this is the easiest way of achieving their desires, they are generally "not evil, just lazy". Practices such as factory farming are widespread not because of some deep-seated desire to cause suffering, but rather because they are the most efficient way of producing meat and other animal source foods. If technologies such as growing meat from cell cultures became more efficient than factory farming, then the desire for efficiency could lead to the elimination of suffering. Similarly, industrialization has reduced the demand for slaves and forced labor as machine labor has become more effective. At the same time, West acknowledges that this is not a knockdown argument against the possibility of massive future suffering, and that the desire for efficiency could still lead to suffering outcomes such as simulated game worlds filled with sentient non-player characters (see section on cruelty-enabling technologies below).

Another argument against net suffering outcomes is offered by Shulman (2012), who discusses the possibility of civilizations spending some nontrivial fraction of their resources constructing computing matter that was optimized for producing maximum pleasure per unit of energy, or for producing maximum suffering per unit of energy. Shulman's argument rests on the assumption that value and disvalue are symmetrical with regard to such optimized states. The amount of pleasure or suffering produced this way could come to dominate any hedonistic utilitarian calculus, and even a weak benevolent bias that led to there being more optimized pleasure than optimized suffering could tip the balance in favor of there being more total happiness. Shulman's

argument thus suggests that net suffering outcomes could be unlikely unless a (non-compassionate) singleton ensures that no optimized happiness is created. However, the possibility of optimized suffering and the chance of e.g. civilizations intentionally creating it as a way of extorting agents that care about suffering reduction, also makes astronomical suffering outcomes more likely.

4.2 Suffering outcome: dystopian scenarios created by non-value-aligned incentives.

Bostrom (2005, 2014) discusses the possibility of technological development and evolutionary and competitive pressures leading to various scenarios where everything of value has been lost, and where the overall value of the world may even be negative. Considering the possibility of a world where most minds are brain uploads doing constant work, Bostrom (2014) points out that we cannot know for sure that happy minds are the most productive under all conditions: it could turn out that anxious or unhappy minds would be more productive. If this were the case, the resulting outcomes could be dystopian indeed:

We seldom put forth full effort. When we do, it is sometimes painful. Imagine running on a treadmill at a steep incline—heart pounding, muscles aching, lungs gasping for air. A glance at the timer: your next break, which will also be your death, is due in 49 years, 3 months, 20 days, 4 hours, 56 minutes, and 12 seconds. You wish you had not been born. (Bostrom 2014, p. 201)

As Bostrom (2014) notes, this kind of a scenario is by no means inevitable; Hanson (2016) argues for a more optimistic outcome, where brain emulations still spend most of their time working, but are generally happy. But even Hanson's argument depends on economic pressures and human well-being happening to coincide: absent such a happy coincidence, he offers no argument for believing that the future will indeed be a happy one.

More generally, Alexander (2014) discusses examples such as tragedies of the commons, Malthusian traps, arms races, and races to the bottom as cases where people are forced to choose between sacrificing some of their values and getting outcompeted. Alexander also notes the existence of changes to the world that nearly everyone would agree to be net improvements - such as every country reducing its military by 50%, with the savings going to infrastructure - which nonetheless do not happen because nobody has the incentive to carry them out. As such, even if the prevention of various kinds of suffering outcomes would be in everyone's interest, the world might nonetheless end up in them if the incentives are sufficiently badly aligned and new technologies enable their creation.

An additional reason for why such dynamics might lead to various suffering outcomes is the so-called Anna Karenina principle (Diamond 1997, Zaneveld et al. 2017), named after the opening line of Tolstoy's novel *Anna Karenina:* "all happy families are all alike; each unhappy family is unhappy in its own way". The general form of the principle is that for a range of endeavors or processes, from animal domestication (Diamond 1997) to the stability of animal microbiomes (Zaneveld et al. 2017), there are many different factors that all need to go right, with even a single mismatch being liable to cause failure.

Within the domain of psychology, Baumeister et al. (2001) review a range of research areas to argue that "bad is stronger than good": while sufficiently many good events can overcome the effects of bad experiences, bad experiences have a bigger effect on the mind than good ones do. The effect of positive changes to well-being also tends to decline faster than the impact of negative changes: on average, people's well-being suffers and never fully recovers from events such as disability, widowhood, and divorce, whereas the improved well-being that results from events such as marriage or a job change dissipates almost completely given enough time (Lyubomirsky 2010).

To recap, various evolutionary and game-theoretical forces may push civilization in directions that are effectively random, random changes are likely to bad for the things that humans value, and the effects of bad events are likely to linger disproportionately on the human psyche. Putting these considerations together suggests (though does not guarantee) that freewheeling development could eventually come to produce massive amounts of suffering.

A possible counter-argument is that people are often more happy than their conditions might suggest. For example, as a widely-reported finding, while the life satisfaction reported by people living in bad conditions in slums is lower than that of people living in more affluent conditions, it is still higher than one might intuitively expect, and the slum-dwellers report being satisfied with many aspects of their life (Biswas-Diener & Diener 2001). In part, this is explained by fact that despite the poor conditions, people living in the slums still report many things that bring them pleasure: a mother who has lost two daughters reports getting joy from her surviving son, is glad that the son will soon receive a job at a bakery, and is glad about her marriage to her husband and feels that her daily prayer is important (Biswas-Diener & Diener 2001).

However, a proper evaluation of this research is complicated: "suffering" might be conceptualized as best corresponding to negative feelings, which are a separate component from cognitively evaluated life satisfaction (Lukas, Diener & Suh 1996), with the above slum-dweller study focusing mainly on life satisfaction. In general, life satisfaction is associated with material prosperity, while positive and negative feelings are associated with psychological needs such as autonomy, respect, and the ability to be able count on others in an emergency (Diener et al. 2010). A proper review of the literature and an analysis of how to interpret the research in terms of suffering risks is beyond the scope of this paper.

4.3 Suffering outcome: cruelty-enabling technologies.

Better technology may enable people to better engage in cruel and actively sadistic pursuits. While active sadism and desire to hurt others may be a relatively rare occurrence in contemporary society, public cruelty has been a form of entertainment in many societies, ranging from the Roman practice of involuntary gladiator fights to animal cruelty in the Middle Ages. Even in contemporary society, there are widespread sentiments that people such as criminals should be severely punished in ways which inflict considerable suffering (part of the Roman gladiators were convicted criminals).

Contemporary society also contains various individuals who are motivated by the desire to hurt others (Torres 2016, 2017a, 2017b, chap 4.), even to the point of sacrificing their own lives in the process. For example, Eric Harris, one of the two shooters of the Columbine High School Massacre, wrote extensively about his desire to rape and torture people, fantasized about tricking women into thinking that they were safe so that he could then hurt them, and wanted the freedom to be able to kill and rape without consequences (Langman 2015). While mass shooters tend to be lone individuals, there have existed more organized groups who seem to have given their members the liberty to act on similar motivations (Torres 2017a), such as the Aum Shinrikyo cult, where dissent or even just "impure thoughts" were punished by rituals amounting to torture and defectors "routinely kidnapped, tortured, imprisoned in cargo crates, subjected to electro shock, drugged in the Astral Hospital or killed outright" (Flannery 2016).

While most contemporary societies reject the idea of cruelty as entertainment, civilizations could eventually emerge in which such practices were again acceptable. Assuming advanced technology, this could take the form of keeping criminals and other undesirables alive indefinitely while subjecting them to eternal torture[8], slaves kept for the purpose of sadistic actions who could be healed of any damage inflicted to them (one fictional illustration of such a scenario recently received widespread popularity as the TV series *Westworld*)[9], or even something like vast dystopian simulations of fantasy warfare inhabited by sentient "non-player characters", to serve as the location of massive multiplayer online games which people may play in as super-powered "heroes".

Particularly in the latter scenarios, the amount of sentient minds in such conditions could be many times larger than the civilization's other population. In contemporary computer games, it is normal for the player to kill thousands of computer-controlled opponents during the game, suggesting that a large-scale game in which a sizeable part of the population participated might instantiate very large numbers of non-player characters per player, existing only to be hurt for the pleasure of the players.

5 Suffering outcomes that may be caused by superintelligence[10]

In the previous section, we discussed possible suffering outcomes that might be realized without a singleton that could prevent them from occurring, and suggested that an appropriately-programmed superintelligence is currently the most likely candidate for forming such a singleton. However, an inappropriately programmed superintelligence could also cause suffering outcomes; we will now turn to this topic.

Superintelligence is related to three categories of suffering risk: *suffering subroutines* (Tomasik 2017), *mind crime* (Bostrom 2014) and *flawed realization* (Bostrom 2013).

5.1 Suffering subroutines

Humans have evolved to be capable of suffering, and while the question of which other animals are conscious or capable of suffering is controversial, pain analogues are present in a wide variety of animals. The U.S. National Research Council's Committee on Recognition and Alleviation of Pain in Laboratory Animals (2004) argues that, based on the state of existing evidence, at least all vertebrates should be considered capable of experiencing pain.

Pain seems to have evolved because it has a functional purpose in guiding behavior: evolution having found it suggests that pain might be the simplest solution for achieving its purpose. A superintelligence which was building subagents, such as worker robots or disembodied cognitive agents, might then also construct them in such a way that they were capable of feeling pain - and thus possibly suffering (Metzinger 2015) - if that was the most efficient way of making them behave in a way that achieved the superintelligence's goals.

Humans have also evolved to experience empathy towards each other, but the evolutionary reasons which cause humans to have empathy (Singer 1981) may not be relevant for a superintelligent singleton which had no game-theoretical reason to empathize with others. In such a case, a superintelligence which had no disincentive to create suffering but did have an incentive to create whatever furthered its goals, could create vast populations of agents which sometimes suffered while carrying out the superintelligence's goals. Because of the ruling superintelligence's indifference towards suffering,

[8] Fictional depictions include Ellison (1967) and Ryding (no date); note that both stories contain very disturbing imagery. A third depiction was in the "White Christmas" episode of the TV series *Black Mirror*, which included a killer placed in solitary confinement for thousands of years while having to listen to a Christmas song on an endless loop.

[9] Another fictional depiction includes Gentle (2004); the warning for disturbing graphic imagery very much applies.

[10] This section reprints material that has previously appeared in a work by one of the authors (Gloor 2016), but has not been formally published before.

the amount of suffering experienced by this population could be vastly higher than it would be in e.g. an advanced human civilization, where humans had an interest in helping out their fellow humans.

Depending on the functional purpose of positive mental states such as happiness, the subagents might or might not be built to experience them. For example, Fredrickson (1998) suggests that positive and negative emotions have differing functions. Negative emotions bias an individual's thoughts and actions towards some relatively specific response that has been evolutionarily adaptive: fear causes an urge to escape, anger causes an urge to attack, disgust an urge to be rid of the disgusting thing, and so on. In contrast, positive emotions bias thought-action tendencies in a much less specific direction. For example, joy creates an urge to play and be playful, but "play" includes a very wide range of behaviors, including physical, social, intellectual, and artistic play. All of these behaviors have the effect of developing the individual's skills in whatever the domain. The overall effect of experiencing positive emotions is to build an individual's resources - be those resources physical, intellectual, or social.

To the extent that this hypothesis were true, a superintelligence might design its subagents in such a way that they had pre-determined response patterns for undesirable situations, so exhibited negative emotions. However, if it was constructing a kind of a command economy in which it desired to remain in control, it might not put a high value on any subagent accumulating individual resources. Intellectual resources would be valued to the extent that they contributed to the subagent doing its job, but physical and social resources could be irrelevant, if the subagents were provided with whatever resources necessary for doing their tasks. In such a case, the end result could be a world whose inhabitants experienced very little if any in the way of positive emotions, but did experience negative emotions. This could qualify as any one of the suffering outcomes we've considered (astronomical, net, pan-generational net).

A major question mark with regard to suffering subroutines are the requirements for consciousness (Muehlhauser 2017) and suffering (Metzinger 2016, Tomasik 2017). The simpler the algorithms that can suffer, the more likely it is that an entity with no regard for minimizing it would happen to instantiate large numbers of them. If suffering has narrow requirements such as a specific kind of self-model (Metzinger 2016), then suffering subroutines may become less common.

Below are some pathways that could lead to the instantiation of large numbers of suffering subroutines (Gloor 2016):

Anthropocentrism. If the superintelligence had been programmed to only care about humans, or by minds which were sufficiently human-like by some criteria, then it could end up being indifferent to the suffering of any other minds, including subroutines.

Indifference. If attempts to align the superintelligence with human values failed, it might not put any intrinsic value on avoiding suffering, so it may create large numbers of suffering subroutines.

Uncooperativeness. The superintelligence's goal is something like classical utilitarianism, with no additional regards for cooperating with other value systems. As previously discussed, classical utilitarianism would prefer to avoid suffering, all else being equal. However, this concern could be overridden by opportunity costs. For example, Bostrom (2003a) suggests that every *second* of delayed space colonization corresponds to a loss equal to 10^14 potential lives. A classical utilitarian superintelligence that took this estimate literally might choose to build colonization robots that used suffering subroutines, if this was the easiest way and developing alternative cognitive architectures capable of doing the job would take more time.

5.2 Mind crime

A superintelligence might run simulations of sentient beings for a variety of purposes. Bostrom (2014, p. 152) discusses the specific possibility of an AI creating simulations of human beings which were detailed enough to be conscious. These simulations could then be placed in a variety of situations in order to study things such as human psychology and sociology, and destroyed afterwards.

The AI could also run simulations that modeled the evolutionary history of life on Earth, to obtain various kinds of scientific information or to help estimate the likely location of the "Great Filter" (Hanson 1998) and whether it should expect to encounter other intelligent civilizations. This could repeat the wild-animal suffering (Tomasik 2015, Dorado 2015) experienced in Earth's evolutionary history. The AI could also create and mistreat, or threaten to mistreat, various minds as a way to blackmail other agents.

As it is possible that minds in simulations could one day compose the majority of all existing minds (Bostrom 2003b), and that with sufficient technology there could be astronomical numbers of them, then depending on the nature of the simulations and the net amount of happiness and suffering, mind crime could possibly lead to any one of the three suffering outcomes.

Below are some pathways that could lead to mind crime (Gloor 2016):

Anthropocentrism. Again, if the superintelligence had been programmed to only care about humans, or about minds which were sufficiently human-like by some criteria, then it could be indifferent to the suffering experienced by non-humans in its simulations.

Indifference. If attempts to align the superintelligence with human values failed, it might not put any intrinsic value on avoiding suffering, so it may create large numbers of simulations with sentient minds if that furthered its objectives.

Extortion. The superintelligence comes into conflict with another actor that disvalues suffering, so the superintelligence instantiates large numbers of suffering minds as a way of extorting the other entity.

Libertarianism regarding computations: the creators of the first superintelligence instruct the AI to give every human alive at the time control of a planet or galaxy,

with no additional rules to govern what goes on within those territories. This would practically guarantee that some humans would use this opportunity for inflicting widespread cruelty (see the previous section).

5.3 Flawed realization

A superintelligence with human-aligned values might aim to convert the resources in its reach into clusters of utopia, and seek to colonize the universe in order to maximize the value of the world (Bostrom 2003a), filling the universe with new minds and valuable experiences and resources. At the same time, if the superintelligence had the wrong goals, this could result in a universe filled by vast amounts of *disvalue*.

While some mistakes in value loading may result in a superintelligence whose goal is completely unlike what people value, certain mistakes could result in *flawed realization* (Bostrom 2013). In this outcome, the superintelligence's goal gets human values *mostly* right, in the sense of sharing many similarities with what we value, but also contains a flaw that drastically changes the intended outcome[11].

For example, value extrapolation (Yudkowsky 2004) and value learning (Soares 2016, Sotala 2016) approaches attempt to learn human values in order to create a world that is in accordance with those values. There have been occasions in history when circumstances that cause suffering have been defended by appealing to values which seem pointless to modern sensibilities, but which were nonetheless a part of the prevailing values at the time. In Victorian London, the use of anesthesia in childbirth was opposed on the grounds that being under the partial influence of anesthetics may cause "improper" and "lascivious" sexual dreams (Farr 1980), with this being considered more important to avoid than the pain of childbirth.

A flawed value-loading process might give disproportionate weight to historical, existing, or incorrectly extrapolated future values whose realization then becomes more important than the avoidance of suffering. Besides merely considering the avoidance of suffering less important than the enabling of other values, a flawed process might also tap into various human tendencies for endorsing or celebrating cruelty (see the discussion in section 4), or outright glorifying suffering. Small changes to a recipe for utopia may lead to a future with much more suffering than one shaped by a superintelligence whose goals were completely different from ours.

6 How and whether to work on s-risk?

In the previous sections, we have argued for s-risks being severe enough to be worth preventing, and for there to be several plausible routes by which they might be realized. We will now argue for the case that it is possible to productively work on them today, via some of the following recommendations.

Carry out general AI alignment work. Given that it would generally be against the values of most humans for suffering outcomes to be realized, research aimed at aligning AIs with human values (Yudkowsky 2008, Goertzel & Pitt 2012, Bostrom 2014, Sotala 2016, Soares & Fallenstein 2017) seems likely to also reduce the risk of suffering outcomes. If our argument for suffering outcomes being something to avoid is correct, then an aligned superintelligence should also attempt to establish a singleton that would prevent negative suffering outcomes, as well as avoiding the creation of suffering subroutines and mind crime.

In addition to technical approaches to AI alignment, the possibility of suffering risks also tends to make more similar recommendations regarding social and political approaches. For example, Bostrom et al. (2016) note that conditions of *global turbulence* might cause challenges for creating value-aligned AI, such as if pre-existing agreement are not kept to and ill-conceived regulation is enacted in a haste. Previous work has also pointed to the danger of arms races making it harder to keep AI aligned (Shulman 2009, Miller 2012, Armstrong et al. 2013). As the avoidance of suffering outcomes is the joint interest of many different value systems, measures that reduce the risk of arms races and improve the ability of different value systems to shape the world in their desired direction can also help avoid suffering outcomes.

Besides making AIs more aligned in general, some interventions may help avoid negative outcomes - such as suffering outcomes from flawed realization scenarios - in particular. Most of the current alignment research seeks to ensure that the values of any created AIs are aligned with humanity's values to a maximum possible extent, so that the future they create will contain as much positive value as possible. This is a difficult goal: to the extent that humanity's values are complex and fragile (Yudkowsky 2011), successful alignment may require getting a very large amount of details right.

On the other hand, it seems much easier to give AIs goals that merely ensure that they will not create a future with *negative* value by causing suffering outcomes. This suggests an approach of fail-safe methods: safety nets or mechanisms such that, if AI control fails, the outcome will be as good as it gets under the circumstances. Fail-safe methods could include tasking AI with the objective of buying more time to carefully solve goal alignment more generally, or fallback goal functions:

Research fallback goals: Research ways to implement multi-layered goal functions, with a "fallback goal" that kicks in if the implementation of the top layer does not fulfill certain safety criteria. The fallback would be a simpler, less ambitious goal that is less likely to result in bad outcomes. Difficulties would lie in selecting the safety criteria in ways that people with different values could all agree on, and in making sure that the fallback goal gets triggered under the correct circumstances.

[11] One fictional illustration of a flawed utopia is Yudkowsky (2009), though this setting does not seem to contain enormous amounts of suffering.

Care needs to be taken with the selection of the fallback goal, however. If the goal was something like reducing suffering, then in a multipolar (Bostrom 2014) scenario, other superintelligences could have an incentive to create large amounts of suffering in order to coerce the superintelligence with the fallback goal to act in some desired way.

Research ways to clearly separate superintelligence designs from ones that would contribute to suffering risk. Yudkowsky (2017) proposes building potential superintelligences in such a way as to make them widely separated in design space from ones that would cause suffering outcomes. For example, if an AI has a representation of "what humans value" V which it is trying to maximize, then it would only take a small (perhaps accidental) change to turn it into one that maximized -V instead, possibly causing enormous suffering. One proposed way of achieving this is by never trying to explicitly represent complete human values: then, the AI "just doesn't contain the information needed to compute states of the universe that we'd consider worse than death; flipping the sign of the utility function U, or subtracting components from U and then flipping the sign, doesn't identify any state we consider worse than [death]" (Yudkowsky 2017). This would also reduce the risk of suffering being created through another actor which was trying to extort the superintelligence.

Carry out research on suffering risks and the enabling factors of suffering. At this moment, there is only little research to the possibility of risks of astronomical suffering. Two kinds of research would be particularly useful. First, research focused on understanding the biological and algorithmic foundation of suffering (Metzinger 2016) could help understand how likely outcomes such as suffering subroutines would be. Pearce (1995) has argued for the possibility of minds motivated by "gradients of bliss", which would not need to experience any suffering: if minds could be designed in such a manner, that might help avoid suffering outcomes.

Second, research on suffering outcomes in general, to understand how to avoid them. With regard to suffering risks from extortion scenarios, targeted research in economics, game theory or decision theory could be particularly valuable.

Rethink maxipok and maximin. Bostrom (2002, 2013) proposes a "maxipok rule" to act as a rule of thumb when trying to act in the best interest of humanity as a whole:

Maxipok: Maximise the probability of an 'OK outcome', where an OK outcome is any outcome that avoids existential catastrophe.

The considerations in this paper do not necessarily refute the rule as written, especially not since Bostrom defines an "existential catastrophe" to include "the permanent and drastic destruction of its potential for desirable future development", and the realization of suffering outcomes could very well be thought to fall under this definition. However, in practice much of the discourse around the concept of existential risk has focused on the possibility of extinction, so it seems valuable to highlight the fact that "existential catastrophe" does not include only scenarios of zero value, but also scenarios of negative value.

Bostrom (2002, 2013) also briefly discusses the "maximin" principle, "choose the action that has the best worst-case outcome", and rejects this principle as he argues that this entails "choosing the action that has the greatest benefit under the assumption of impending extinction. Maximin thus implies that we ought all to start partying as if there were no tomorrow." (Bostrom 2013, p. 19). However, since a significant contribution to the expected value of AI comes from worse outcomes than extinction, this argument is incorrect. While there may be other reasons to reject maximin, the principle correctly implies choosing the kinds of actions that avoid the worst suffering outcomes and so might not be very dissimilar from maxipok.

7 Acknowledgments

David Althaus, Stuart Armstrong, Tobias Baumann, Max Daniel, Ruairi Donnelly, and two anonymous reviewers provided valuable feedback on this paper.

8 References

[1] Alexander, S. (2014). Meditations on Moloch. *Slate Star Codex*. http://slatestarcodex.com/2014/07/30/meditations-on-moloch/

[2] Armstrong, S., & Sandberg, A. (2013). Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. *Acta Astronautica*, *89*, 1-13.

[3] Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: a model of artificial intelligence development. *AI & Society*, *31*(2), 201-206.

[4] Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323-370.

[5] Benatar, D. (2008). Better never to have been: the harm of coming into existence. Oxford University Press.

[6] Biswas-Diener, R. & Diener, E. (2001). Making the best of a bad situation: Satisfaction in the slums of Calcutta. *Social Indicators Research*, 55, 329-352.

[7] Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, *9*(1).

[8] Bostrom, N. (2003a). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, *15*(3), 308-314.

[9] Bostrom, N. (2003b). Are we living in a computer simulation? *The Philosophical Quarterly*, 53(211), 243-255.

[10] Bostrom, N. (2004). The future of human evolution. In Tandy, C. (ed.) *Death and anti-death: Two hundred years after Kant, fifty years after Turing*, 339-371.

[11] Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15-31.

[12] Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. OUP Oxford.
[13] Bostrom, N., Dafoe, A., & Flynn, C. (2016). Policy Desiderata in the Development of Machine Superintelligence. https://nickbostrom.com/papers/aipolicy.pdf
[14] Diamond, J. (1997). Guns, Germs, and Steel: The Fates of Human Societies. W. W. Norton.
[15] Diener, E., Ng, W., Harter, J. & Arora, R. (2010). Wealth and Happiness Across the World: Material Prosperity Predicts Life Evaluation, Whereas Psychosocial Prosperity Predicts Positive Feeling. *Journal of Personality and Social Psychology, 99(1)*, 52– 61.
[16] Ditto, P. H., Druley, J. A., Moore, K. A., Danks, J. H., & Smucker, W. D. (1996). Fates worse than death: the role of valued life activities in health-state evaluations. *Health Psychology*, *15*(5), 332.
[17] Dorado, D. (2015). Ethical Interventions in the Wild: An Annotated Bibliography. *Relations: Beyond Anthropocentrism, 3*, 219.
[18] Ellison, H. (1967). I Have No Mouth, and I Must Scream. *IF: Worlds of Science Fiction*, March 1967.
[19] Farr, A. D. (1980). Early opposition to obstetric anaesthesia. *Anaesthesia, 35*(9), 896-907.
[20] Fehige, C. (1998). "A Pareto Principle for Possible People" in Fehige, C. & Wessels, U. (eds.) *Preferences.* Berlin: De Gruyter, 509-543.
[21] Food and Agriculture Organization of the United Nations. (2012). FAOSTAT Agriculture – Livestock Primary Dataset: world total of animals slaughtered for meat in 2012. Retrieved January 26, 2016, from http://faostat.fao.org/site/569/DesktopDefault.aspx?PageID=569
[22] Flannery, F. (2016). Understanding Apocalyptic Terrorism: Countering the Radical Mindset. Abingdon, Oxon: Routledge.
[23] Fredrickson, B. L. (1998). What good are positive emotions? *Review of General Psychology, 2*(3), 300.
[24] Gentle, M. (2004). Human Waste. In *Cartomancy*. Gollancz.
[25] Gloor, L. (2016). Suffering-focused AI safety. Foundational Research Institute. https://foundational-research.org/suffering-focused-ai-safety-why-fail-safe-measures-might-be-our-top-intervention/
[26] Gloor, L. (2017). Tranquilism. Foundational Research Institute. https://foundational-research.org/tranquilism/
[27] Gloor, L. & Mannino, A. (2016). The Case for Suffering-Focused Ethics. Foundational Research Institute. https://foundational-research.org/the-case-for-suffering-focused-ethics/
[28] Goertzel, B. & Pitt, J., (2012). Nine ways to bias open-source AGI toward friendliness. *Journal of Evolution and Technology, 22*(1).
[29] Greaves, H (2017). Population axiology. *Philosophy Compass*, 12:e12442. https://doi.org/10.1111/phc3.12442
[30] Greene, J. (2013). Moral tribes: Emotion, Reason, and the gap between us and them. Penguin.
[31] Hanson, R. (1998). The Great Filter - Are We Almost Past It? http://mason.gmu.edu/~rhanson/greatfilter.html
[32] Hanson, R. (2016). The Age of Em: Work, Love, and Life when Robots Rule the Earth. Oxford University Press.
[33] Häggström, O. (2016). Here Be Dragons: Science, Technology and the Future of Humanity. Oxford University Press.
[34] Langman, P. (2015). School Shooters: Understanding High School, College, and Adult Perpetrators. Lanham, MD: Rowman & Littlefield.
[35] Le Guin, U. K. (1973). The ones who walk away from Omelas. In Silverberg, R. (ed.), *New Dimensions 3*.
[36] Lukas, R.E., Diener, E. & Suh, E. (1996). Discriminant Validity of Well-Being Measures. *Journal of Personality and Social Psychology*, 71(3), 616-628.
[37] Lyubomirsky, S. (2010). 11 Hedonic Adaptation to Positive and Negative Experiences. In Folkman, S. & Nathan, P.E. (eds.) *The Oxford Handbook of Stress, Health, and Coping*. Oxford University Press.
[38] Manning, P. (1992) The Slave Trade: The Formal Demography of a Global System. In Klein, M. A., & Hogendorn, J. (eds.) *The Atlantic slave trade: effects on economies, societies and peoples in Africa, the Americas, and Europe*. Duke University Press.
[39] Metzinger, T. (2015). What if they need to suffer? Edge.org response to: What do you think about machines that think? https://www.edge.org/response-detail/26091
[40] Metzinger, T. (2016). Suffering. In Almqvist, K. & Haag, A. (eds.) *The Return of Consciousness*. Stockholm: Axel and Margaret Ax:son Johnson Foundation
[41] Miller, J. D. (2012). Singularity Rising: Surviving and thriving in a smarter, richer, and more dangerous world. BenBella Books, Inc.
[42] Muehlhauser, L. (2017). 2017 Report on Consciousness and Moral Patienthood. *Open Philanthropy Project.* https://www.openphilanthropy.org/2017-report-consciousness-and-moral-patienthood
[43] Narveson, J. (1973). Moral problems of population. *The Monist*, 62-86.
[44] National Research Council's Committee on Recognition and Alleviation of Pain in Laboratory Animals. (2009). Recognition and alleviation of pain in laboratory animals. Washington (DC): National Academies Press .
[45] Nussbaum, M. (1997) Capabilities and Human Rights, 66 Fordham L. Rev. 273.
[46] Ord, T. (2013). Why I'm Not a Negative Utilitarian. http://www.amirrorclear.net/academic/ideas/negative-utilitarianism/

[47] Parfit, D. (1991). Equality or Priority? (Department of Philosophy: University of Kansas)
[48] Pearce, D. (1995). The Hedonistic Imperative. https://www.hedweb.com/hedab.htm
[49] Persson, I., & Savulescu, J. (2012). *Unfit for the future: the need for moral enhancement.* Oxford University Press.
[50] Rachels, S (1998). Counterexamples to the transitivity of better than. *Australasian Journal of Philosophy,* 76(1):71-83.
[51] Regan, T (1980). Utilitarianism, Vegetarianism, and Animal Rights. *Philosophy and Public Affairs*, 9(4):305-324.
[52] Rubin, E. B., Buehler, A. E., & Halpern, S. D. (2016). States worse than death among hospitalized patients with serious illnesses. *JAMA Internal Medicine, 176*(10), 1557-1559.
[53] Ryding, D. (no date). Yes, Jolonah, There Is A Hell. *Orion's Arm.* http://www.orionsarm.com/page/233
[54] Sayah, F. A., Mladenovic, A., Gaebel, K., Xie, F., & Johnson, J. A. (2015). How dead is dead? Qualitative findings from participants of combined traditional and lead-time time trade-off valuations. *Quality of Life Research, 25*(1), 35-43.
[55] Shulman, C. (2012). Are pain and pleasure equally energy-efficient? **Reflective Disequilibrium**. https://reflectivedisequilibrium.blogspot.fi/2012/03/are-pain-and-pleasure-equally-energy.html
[56] Shulman, C., & Armstrong, S. (2009). Arms control and intelligence explosions. In *7th European Conference on Computing and Philosophy (ECAP)*, Bellaterra, Spain, July (pp. 2-4).
[57] Singer, P. (1981/2011). The Expanding Circle: Ethics, Evolution, and Moral Progress. Princeton University Press.
[58] Singer, P. (1993). *Practical Ethics*, second edition. Cambridge University Press.
[59] Soares, N., & Fallenstein, B. (2017). Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In Callaghan et al. (eds.) *The Technological Singularity - Managing the Journey* (pp. 103-125). Springer Berlin Heidelberg.
[60] Soares, N. (2016). The Value Learning Problem. *2nd International Workshop on AI and Ethics, AAAI-2016.* Phoenix, Arizona.
[61] Sotala, K., & Yampolskiy, R. V. (2015). Responses to Catastrophic AGI Risk: A Survey. *Physica Scripta, 90*(1), 018001.
[62] Sotala, K. (2016). Defining Human Values for Value Learners. *2nd International Workshop on AI and Ethics, AAAI-2016.* Phoenix, Arizona.
[63] Tomasik, B. (2015). The importance of wild-animal suffering. *Relations: Beyond Anthropocentrism, 3*, 133.
[64] Tomasik, B. (2017). What Are Suffering Subroutines? http://reducing-suffering.org/what-are-suffering-subroutines/
[65] Torres, P. (2016).Agential Risks: A Comprehensive Introduction. Journal of Evolution and Technology, 26(2), pp. 31-47.
[66] Torres, P. (2017a). Who Would Destroy the World? Omnicidal Agents and Related Phenomena. Pre-publication draft: https://docs.wixstatic.com/ugd/d9aaad_b18ce62c32be44ddbf64268fc295fdc0.pdf
[67] Torres, P. (2017b). Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks. Durham, North Carolina: Pitchstone Publishing.
[68] West, B. (2017). An Argument for Why the Future May Be Good. *Effective Altruism Forum*, http://effective-altruism.com/ea/1cl/an_argument_for_why_the_future_may_be_good/.
[69] World Food Programme. (2017). Zero Hunger. http://www1.wfp.org/zero-hunger
[70] Yudkowsky, E. (2004). Coherent extrapolated volition. Singularity Institute for Artificial Intelligence. https://intelligence.org/files/CEV.pdf
[71] Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In Bostrom, N. & Ćirković, M.M. (eds.) *Global Catastrophic Risks.* New York: Oxford University Press.
[72] Yudkowsky, E. (2009). Failed Utopia #4-2. *Less Wrong.* http://lesswrong.com/lw/xu/failed_utopia_42/
[73] Yudkowsky, E. (2011). Complex Value Systems are Required to Realize Valuable Futures. Machine Intelligence Research Institute. https://intelligence.org/files/ComplexValues.pdf
[74] Yudkowsky, E. (2017). Separation from hyperexistential risk. *Arbital.* Retrieved December 11, 2017, from https://arbital.com/p/hyperexistential_separation/
[75] Zaneveld, J. R., McMinds, R., & Vega, T. R. (2017). Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nature Microbiology, 2*, 17121.

Artificial Intelligence in Life Extension: from Deep Learning to Superintelligence

Mikhail Batin and Alexey Turchin
Science for Life Extension Foundation, Prospect Mira 124-15, Moscow, Russia
E-mail: alexeiturchin@gmail.com, http://scienceagainstaging.com/

Sergey Markov
ActiveBusinessCollection, Russia, Moscow, d.19 ul. Vavilova, Moscow 117997, Russia
E-mail: sergei.markoff@gmail.com, https://activebc.ru/

Alisa Zhila
IBM Watson, IBM Corporation, 1 New Orchard Road, Armonk, NY 10504-1722, USA
E-mail: alisa.zhila@gmail.com, https://www.ibm.com/watson/

David Denkenberger
Global Catastrophic Risk Institute; Tennessee State University
Alliance to Feed the Earth in Disasters; 3500 John A Merritt Blvd, Nashville, TN 37209, USA
E-mail: david.denkenberger@gmail.com, http://allfed.info/

Keywords: artificial intelligence, life extension, aging, geroprotectors, biomarkers, machine learning

Received: August 31, 2017

In this paper we focus on the most efficacious AI applications for life extension and anti-aging at three expected stages of AI development: narrow AI, AGI and superintelligence. First, we overview the existing research and commercial work performed by a select number of startups and academic projects. We find that at the current stage of "narrow" AI, the most promising areas for life extension are geroprotector-combination discovery, detection of aging biomarkers, and personalized anti-aging therapy. These advances could help currently living people reach longevity escape velocity and survive until more advanced AI appears. When AI comes close to human level, the main contribution to life extension will come from AI integration with humans through brain-computer interfaces, integrated AI assistants capable of autonomously diagnosing and treating health issues, and cyber systems embedded into human bodies. Lastly, we speculate about the more remote future, when AI reaches the level of superintelligence and such life-extension methods as uploading human minds and creating nanotechnological bodies may become possible, thus lowering the probability of human death close to zero. We suggest that medical AI based superintelligence could be safer than, say, military AI, as it may help humans to evolve into part of the future superintelligence via brain augmentation, uploading, and a network of self-improving humans. Medical AI's value system is focused on human benefit.

Povzetek: Prispevek opisuje najbolj učinkovite aplikacije umetne inteligence za podaljšanje življenjske in delovne dobe od klasičnega strojnega učenja do superinteligence.

1 Introduction

The 2010s have shown a rapidly growing interest in Artificial Intelligence (AI) technologies [63]. In recent years, AI has appeared in top scientific news sources, in stories that have demonstrated that AI is "smarter" than humans when it comes to playing a number of boardgames [89] and word games [61], thus revealing that AI is approaching a revolutionary point in its development.

Investments in AI-related projects have increased dramatically in the last few years. Global AI startup financing reached US$5 billion in 2016 [76]. The current market of AI in medicine is estimated at US$1.1 billion and is expected to grow to US$9.1 billion in the next decade [118]. Major IT companies including Google, Facebook, IBM, Intel, and Microsoft nearly simultaneously established biomedical subdivisions because their leadership sees great potential for AI in healthcare. Based on the current rate of development, it is probable that AI will become a revolutionary technology in healthcare in the upcoming decades.

AI has the potential to have the greatest impact on the human life span through life-extension technologies, but the means are underexplored. In this article we investigate which AI technologies in healthcare are likely to provide the best results in the quest for increased life expectancy. There is a great number of publications about the practical applications of existing AI in medicine and healthcare. A recent review performed by Ching, et al. [24] describes opportunities and obstacles for the applications of deep learning in medicine. Unlike their review, ours concentrates on expected applications of different stages of AI

development to fight the main cause of death in humans, aging. We demonstrate how gradual evolution of AI in medicine will result in medically oriented beneficial superintelligence able to produce indefinite life extension.

The considered time span also distinguishes this work from other analyses of benevolent AI, such as [16] and [58], which immediately jump to the stage of superintelligence, when AI will, by definition, be able to solve most or all of our problems. As AI is constantly evolving, we should determine how to use it most efficiently during each stage of its development and look at the period between now and superintelligence. Only by doing this will we be able to achieve the longest possible life extension for currently-living human beings.

In this article we outline a path for the application of AI to life extension that yields increasing gains at each step. We show that analysis of aging biomarkers and geroprotectors with the use of narrow AI will make the largest impact on human life expectancy with a relatively small investment. We also show how an increasing amount of an individual's healthcare data collected via wearable devices ("wearables") will feed the data-crunching ability of AI and provide constant personalized monitoring of that individual's health on ever-deeper levels, thus preventing illness at earlier stages as well as repairing age-related damage. We also demonstrate how AI-powered robotics will gradually become inner parts of the human body, resulting in *cyborgization* and high survivability. Our final point of interest is integration of AI with the human brain via neuroimplants to enable mind uploading. See table 1 for an outline of the expected evolution of the application of medical AI in life extension.

The growth of AI's ability for independent research will be increasingly helpful in finding new technologies to lower human mortality until AI reaches the stage of self-improvement. We expect that the development of medical AI will at least partly offset the existential AI risk [16] via intrinsic orientation of medical AI on human benefit and AI's closer integration with humans via brain implants (see section 7.2).

This article is conceptually similar to the report on the expected development of military AI [28], in which the same three levels of the future of AI are considered. The idea that AI will help us to make large gains in life expectancy has been explored in works of futurists Ray Kurzweil [58] and Robert A. Freitas Jr. [36], among others.

This paper is structured as follows. In section 2, we review the expected progress in AI, the levels of development of AI, and the predicted timeline for the corresponding advances. In section 3, we review the current applications of AI to life extension, as developed by select startups and academic projects. Prospective near-future applications of AI to life extension and anti-aging are outlined in section 4, which covers research that is yet to be transferred from academia to the life-extension industry. The expected effect of artificial general intelligence (AGI) on life extension and applications that it will enable are discussed in section 5. The more distant future of AI, including superintelligence and its effect on life expectancy, is outlined in section 6. In section 7, we conclude our overview with a discussion of the best strategies for using AI to maximize the life span of the currently living generation.

Epoch / Sphere	Narrow AI and machine learning	Human-level AI	Superintelligence
Safety and healthcare ecosystem	Patient organizations; Digital medicine	Global safety system	Merge of humanity and AI; Indefinite life extension
Integration of AI and human brain	Intellectual avatar; Digital immortality	Exocortex; Upgrade net	Uploading; Neuroweb
Integration of AI and human body	Wearables	Cyborgization; Microrobots	Nanomedicine
Scientific research in life extension	Geroprotectors; Biomarkers of aging;	Artificial scientists; Full control of genome	Full control of biology

Table 1: Expected evolution of medical AI in life extension.

2 AI development in the twenty-first century

2.1 AI development pace

Predictions about the development of AI have been complicated by AI "winters," periods of decline in funding and enthusiasm due to the lack of breakthroughs.

Despite past "winters," the advancement of AI technologies has skyrocketed in recent years. We are living in a very exciting moment, considering the overall rise in enthusiasm for AI. According to one survey [16], a majority of scientists believe that human-level AI, then superintelligence, will be achieved before the end of the twenty-first century. The current moment (2016–2017), is a period of accelerated AI development, fueled partly by the hype surrounding neural networks and machine

learning. Dozens of startups are working to develop AGI, and they are attracting substantial funding. Achievements in the development of AI are doubling every year in such areas as complexity in text understanding, speech and visual recognition, and natural language conversation [33].

If we extrapolate current trends in the performance and capacity of neural networks, infrahuman (that is able to most things that can do ordinary human being and may work as a robotic brain; but some complex creative activity is still beyond its abilities). AI could be achieved as soon as the 2020s [93].

A recent, large poll of AI scientists [41] shows that AI is expected to be able to master human language around 2026 and, with 50 percent confidence, that machines will exceed humans in every task by 2062.

If AGI appears soon enough, its impact will overshadow that of the slower, decade-long research in geroprotectors described below, and thus make them obsolete even before their fruition, as AGI will provide better solutions. Yet we cannot rely on the early AGI scenario, as AI prediction is known to be difficult.

In any case, two possible scenarios are:

- AGI will be achieved in the coming two decades;
- AGI will be achieved by the end of the twenty-first century.

There is a big practical difference between these two scenarios. In the first case, the majority of people living today will be able to use AI for almost indefinite life extension. In the second case, most currently living people will be able to enjoy the benefits of AGI only if a huge effort is made to take advantage of all intermediate life-extension technologies to help the current population survive to see AGI achieved.

Aubrey de Grey named the situation of improving life expectancy rate equal to the passage of time "longevity escape velocity" [4]. The result would be indefinite life expectancy (ignoring accidents, global catastrophes, etc.). In this paper we show that AI is the main "game changer" that will help currently living people reach longevity escape velocity, as its effects over time will outweigh other known means of life extension. AI is the most rapidly developing technology, and it affects and accelerates the development of all other life-extension technologies.

The exponential growth of AI, which is now doubling with a period of one year, according to [33], will potentially be able to compensate for the exponential growth of the probability of human death because of aging, which doubles every seven years [37], but there is large lag of implementation of medical AI technology. However, it is possible that AI growth will slow down, as it happened several times before during AI winters, and will be sigmoidal.

In [15], Nick Bostrom shows that each day of delay in the achievement of superintelligent AI, which would reverse aging, costs 100 thousand human lives.

The pace of the AI progress is very uncertain but for the purpose of this article, we are going to talk about stages of AI development in a way that is agnostic to timelines.

2.2 The three levels of the future of AI development

In this section we clarify and enhance the classification of the levels of the prospective AI. These levels are often mixed in AI discussion, which leads to confusion.

Narrow AI (weak AI) is the level of a computer program that achieves above-human performance in a specific, narrow task [16]. For example, the tasks of MRI scan recognition and facial recognition require two differently trained systems, although the underlying learning mechanism may be the same. Most existing AI systems are considered narrow AI. The number of such programs is growing rapidly due to the success of machine learning and neural networks.

The difference between narrow AI and conventional computer programs is the ability of the former to learn. Autonomous cars employ a good example narrow AI. Such AI systems do not have full human capacity, particularly in generalization.

Additionally, the majority of contemporary AI systems need ongoing human supervision.

AGI (human-level AI) is AI at the level of human intelligence in many areas. For example, there would likely be communication in natural language, understanding the context of most situations, as well as performing most of the intellectual tasks that humans are able to perform.

Philosophical questions about the possibility of consciousness in AI are outside the scope of this pragmatic definition. Ability to self-improve is an obvious consequence of this level of AI development. As a result, according to Nick Bostrom [16], an era of human-level AI will be brief, as AGI with self-improving abilities will soon evolve superintelligence. Robin Hanson [45] adheres to the view that computer models—emulations—of the human brain will dominate in the future.

Superintelligence is the level at which AI will supersede humans in all aspects, overtaking the intelligence of the entirety of human civilization. It will be able to govern the world, make scientific discoveries, launch space exploration, and create accurate simulations of the human past. Bostrom [16], Yampolskiy [113], Yudkowsky [114], and many other scientists expect its eventual appearance.

3 The current applications of AI in healthcare and medical research

3.1 Growth of investments in healthcare AI

In 2014–16 the giants of the IT industry announced the launch of biotechnology and life-extension projects based on machine-learning techniques. Among those projects are Google's Calico, focusing on anti-aging; Facebook's Chan Zuckerberg Biohub, searching for drugs for all diseases and creating an atlas of cells for this task; IBM's Watson Health, targeting healthcare in

general; Intel's large biotech section [52]; Microsoft's innovative cloud computations for new drug discovery; and Apple's platform for wearables and software for health monitoring.

Not only big business invests in healthcare research and development; many startups are also making great strides. It is estimated that in 2016, there were 106 startups that used AI in various areas of healthcare. The number of mergers and acquisitions in healthcare AI grew from less than 20 in 2012 to nearly 70 in 2016 [51].

Many startups promise almost unbelievable feats. A collection of press releases for such companies comprises hundreds of pages of breathtaking announcements and lengthy enumerations, but most projects vanish within a few years as the survival rate of startups is low [38]. In order to attract investors, promises are often exaggerated. However, these promises may be used to measure general trends and expectations in the industry.

We can expect investment in AI to grow in the next years if a new AI winter does not occur. The healthcare sector is the largest potential source of funding for AI [11], as it is still a "deficit market" due to a large, unmet demand for better health.

3.2 AI in medical research

Even in scientific research, it is necessary to distinguish between "advertising" statements that often exaggerate achievements and real practical achievements. As to the former, in 2009 it was stated that a robot called Adam was able to formulate hypotheses and conduct experiments on the yeast genome [95]. But there were no subsequent publications on this device.

On the other hand, robots have indeed made substantial contributions to the automation of laboratory studies. For instance, robotic manipulators have automated repetitive operations with test tubes [13].

Among the recent practical applications of AI is the use of artificial neural networks for visual recognition of brain scans, including reconstruction of the relationships between biological neurons in brain connections [25].

Several companies are using AI to accelerate their research:

Gero (formerly known as Quantum Pharmaceuticals) employs the methods of physical kinetics and the modern theory of dynamical systems to model aging processes in complex biological regulatory networks [27] aiming to develop novel anti-aging therapies. To control the health effects of the future drugs Gero team has applied a deep convolutional neural network (CNN) to time series representing human locomotor activity from wearable devices, which allowed to produce a digital biomarker of aging [28]. This biomarker now serves as the scientific basis for Gero lifespan/health risks estimation app[1] and could be used as a metrics of health outcomes for wellness and life insurance industries.

Deep Genomics is working on a system that will allow studying, predicting, and interpreting how genetic variations change important cellular processes such as transcription, splicing, and so on. [119].

Atomwise aims to reduce the cost of new-drug development through the use of a supercomputer and a database of molecular structures to predict which versions of a potential drug will work and which will not. [120].

There are many other companies and scientific groups that use AI to accelerate their medical research, and competition is fierce. Not all of them will survive.

3.3 AI in diagnosis

Claims that AI has outperformed humans in various narrow areas of healthcare have appeared since the 1980s [18]. In the early days, such claims mostly referred to expert systems that were popular at the time. It was difficult to translate such success into wider practice, though—and this scaling issue has plagued AI research from the beginning.

Yet humans are not much better. It was found that in 88% of cases a second opinion gives a different diagnosis [104]. Of course, this estimate may be unrepresentative, as only uncertain cases require additional evaluation, yet it demonstrates uncertainty in human diagnostics.

In April 2016, it was stressed by Mark Zuckerberg that machine learning helps to make diagnosis more accurate, inexpensive, and, perhaps most important, quick [46]. For example, an app that tracks changes in moles based on photos taken with a cell-phone camera can replace expensive visits to a doctor. This software, **Total Body Photography**, analyzes photos of moles in comparison with images of 50 million malignant moles using Israeli image recognition technology [88].

AI will be able to simulate biological processes in the human body and use the resulting models for prediction and diagnosis. This is done by using "big data"—that is, by combining a vast amount of data collected from wearables with the extensive data accumulated in previous medical practice. In 2016, IBM bought several corporations that had extensive data on an enormous number of patients. One of these, **Truven,** which alone has hundreds of millions of medical records, has been bought for US$2.6 billion [26].

AI is also working with text and natural language, which helps to handle scientific papers, medical records, and patient complaints, but it still has considerable difficulty understanding human language [7].

IBM Watson for Oncology is a cognitive-computing system that can answer questions formulated in a natural language (that is, in a human language). It has access to various sources of data: encyclopedias, databases of scientific articles, and knowledge ontologies. Thanks to its huge computing power and preprocessed sources, it can give accurate answers questions it is asked.

Since 2013, IBM Watson has been used at the Memorial Sloan Kettering Cancer Center to facilitate decision-making about treatment of patients with lung cancer. Its database is constantly updated with new disease records.

[1] https://itunes.apple.com/us/app/gero-lifespan/id1222911907

IBM Medical Sieve "is an ambitious long-term exploratory grand challenge project to build a next generation cognitive assistant with advanced multimodal analytics, clinical knowledge and reasoning capabilities that is qualified to assist in clinical decision making in radiology and cardiology" [50].

Google DeepMind (DM) Health is a Google DeepMind subproject that applies AI technology to healthcare [29]. In collaboration with the University College London Hospital, DM will be involved in an algorithm-development project for automated distinguishing between healthy and cancerous tissues in the head and neck area.

Babylon Health (iOS, Android) is a mobile application that allows a user to have an online consultation with a British or Irish doctor [5].

Turbine.ai is a team of scientists that formulate personalized methods of treatment for any type of cancer based on AI. [98].

Insilico Medicine is another startup working on the implementation of deep learning in drug discovery.

3.4 AI in bioinformatics and modeling of living organisms

Often artificial intelligence is thought of as something that people have not experienced yet, and when it becomes familiar and accessible, it stops being perceived as AI and is perceived more as a mere "computational method." A set of such computational methods in biology is called bioinformatics. The field of bioinformatics consists of analysis of the genome, its changes, genome linking to proteins, conformation of proteins, and the evolution of living organisms in general.

The next step in the development of bioinformatics is simulation of living organisms. To make this happen, an entity needs data on cellular processes, huge computing power, and adequate biological models.

One of the first computer models of a living cell was created at Stanford in 2012 [54]. It was the simplest mycoplasma, with only 525 genes. However, Craig Venter, who was working with the same mycoplasma in 2015, recognized that the functions of some 90 genes were unknown, and therefore the completeness of the model is in question [49]. Venter managed to create a viable synthetic organism (*Mycoplasma mycoides* JCVI-syn3.0), whose genome consists of 473 genes, but 149 of them were not fully understood [117].

Cell modeling cannot always be accurate, as it has many levels of uncertainty, starting from the quantum level and protein folding, Brownian motion, and so on. Quantum computers may help with protein-folding modeling in the future.

So far, the most advanced simulation of a multicellular organism has been carried out on the *Caenorhabditis elegans* worm [77]. The simulation includes a model of its "brain," which consists of 302 neurons, and the *connectome* of which has been known for a long time [110]. Some of its functions have been put into the model, but full, correct modeling of its behavior has not been achieved yet.

Modeling of a human cell is much more complex than modeling of a mycoplasma cell because it includes up to 40 times more genes, but such a model will allow medication testing through computer simulation. It will also allow preclinical testing on a variety of substances as well as determining the positive effects of a particular medication positive and how it works. Any divergence from an experiment will contribute to the model's improvement. For now, "organ-on-a-chip" works as a proxy for *in vitro* and *in silico* research [80].

The next stage of this approach will be the modeling of a particular human organs and then full body based on its genome, epigenome, and data from medical analysis. Such a model will enable precise calculation and definition of a medical intervention when required [10].

Big companies are interested in cell modeling as well. Chan Zuckerberg Biohub, for instance, has begun work on the atlas of all human cells [121].

3.5 Merging computational biology, cell programming, and AI

Cell programming is akin to bionanorobotics: making a cell perform more and more complex tasks, including calculations, guided moving, and most importantly, protein creation in specified locations. One of the main applications of the technology is drug delivery to fight cancer.

However, to program cells, one needs to process enormous amount of data about their DNA networks. This is where AI and machine learning come in.

The Cellos project [47], which was presented to the public in 2016, performs DNA-design automation for new living organisms. It can calculate (and then synthesize) a DNA sequence that corresponds to a certain function carried out for specified cell types. Boolean logic (commands such as "AND" and "OR") can be used in this function.

Molecula Maxima [69] is a similar platform, which is positioned as a programming language for genetic engineering.

It is worth mentioning **DNA origami** technology [6], which allows the construction of different microscopic mechanisms from DNA. It is enabled through a very powerful system of computer-aided design that can decompose a designed project into its component elements (blocks), and then write the DNA code that will guide self-assembly into a predetermined shape.

3.6 AI, wearables, and big data

There are hundreds of different medically oriented wearables on the market, the explosion of which began several years ago with fitness trackers such as **Fitbit**. Other wearables include professional medical monitoring devices, such as devices that track heart abnormalities.

The **BioStampRC** sensor [122] is a patch that can be glued to different parts of a body, and it collects various kinds of data and automatically loads them into the cloud.

Similar to wearables are medical implants. One example is an **implanted cardiac defibrillator (ICD)**, which was been used to give an electric shock to restart the heart and save a soccer player on the field [21].

It might be possible to improve the situation by introducing AI trained on large amounts of data in order to define the probabilities of successful ICD therapy for a particular patient in a particular case.

Final Frontier Medical Devices produces devices that can diagnose 90% of emergency situations at home. [109].

Nimb is a wearable ring for requesting emergency help. [123].

Wearables can collect chemical signals from the skin or electrical signals from the brain and heart. The next stage in the development of wearables will involve integrating them more closely with the human body and reducing their size.

Wearables have improved clinical trials by constantly measuring numerous parameters as well as tracking whether drugs have been taken. **AiCure** requires taking a photo of a pill in a patient's mouth [124].

A general trend is that smartphones "absorb" specialized gadget functions. This has happened with fitness trackers, which are currently being replaced by the **Argus** app. Current smartphones can measure blood oxygenation with their camera, replacing a US$50 monitoring gadget with a US$5 app.

Besides the cost savings, the body space limits the number wearables that can be used at one time (setting aside the inconvenience of keeping multiple devices charged and updated). Hence, incorporating all wearables into one device is reasonable. The future universal device will likely combine a smartphone, medical device, and brain-computer interface, and might well take a wearable form such as glasses (**Google Glass**, for example) or a necklace.

Wearables will work together with different safety systems, integrating with infrastructure and optimizing the performance of smart homes [12], self-driving cars, robot police, surveillance, drones, and the "Internet of things," providing a ubiquitous safety and healthcare net. Even toilets can be made "smart," analyzing biological material every time you visit them [91], [116]. Google has already patented a smart bathroom [59].

3.7 The problem of research data verification: blockchain and evidence systems

There is a reproducibility crisis medicine [53]. It is explained by a number of statistical biases as well as fraud and market pressure. Life-extension studies are especially susceptible to fraud, as people are willing to pay for "youth," and it is not easy to make objective measurements in such studies. By being able to work through a large amount of patient data, AI will increase the reliability of results.

Experiment automation, experiment-procedure recording, and the use of blockchain [70] to keep records secure could simplify verification processes and reduce bias and fraud in the field.

4 Prospective applications of AI in aging research

4.1 Fighting aging as the most efficient means for life extension

It is widely understood nowadays that the purpose of general healthcare is not only to treat certain diseases but also to prolong *healthy* human life span.

Different applications of AI in healthcare have different effects on life expectancy. For example, fighting rare diseases or advanced stages of cancer will not yield much increase in total life expectancy over the entire population.

The main causes of death in the US are circulatory diseases (23.1% cardiac deaths, 5.1% stroke deaths), cancer (22.5%), chronic lower respiratory disease (5.6%), and Alzheimer's disease (3.6%). Combined, these conditions cause 59.9% of all deaths in the United States [44]. The probability of these diseases increases exponentially according to the Gompertz law of mortality [66, 67]. More than 75% of all deaths happen to people of 65 years of age or older [40].

As a result, some authors [105], [115] say that aging is the main cause of death and that if we are able to slow the aging process, we will lower the probability of age-related diseases and increase the healthy life span. Experiments show that even simple interventions can slow the aging process and thus delay the onset of deadly diseases in and extend the healthy life span of the *C. elegans* worm [20], mice [66], and rats [87].

These life-extension experiments on animals have involved relatively simple interventions, such as administering long-known drugs (metformin or rapamycin, for example) or restricting caloric intake. Such life-extending drugs are called *geroprotectors* [71].

Unfortunately, studies of the life-extending effects of geroprotectors on humans are scarce, although similar interventions have often been used for other diseases (treating diabetes with metformin, for example), hence proving their safety. Although such studies could have begun long ago, this has not happened, because of a number of social and economic reasons. Naturally, such experiments would require a lot of time (longitudinal experiments take decades) and test groups would need to be large.

Yet there is not the luxury of decades and centuries for classical experiments, as people are dying now, during our lifetime. There is a need to find ways to extend human life—and prove that these inventions work—in a shorter time. A well-recognized way to do this is to find *aging biomarkers* that will track that aging is slowing before all participants of an experiment die.

In short, to slow the aging process, we must find efficient geroprotectors and combinations of geroprotectors; to prove that they work, we need to have independently verified aging biomarkers.

There are many other advanced ideas in the fight against aging, including gene therapy, stem cell research, and Strategies for Engineered Negligible Senescence (SENS) [27]. However, in this section we will limit ourselves to AI-based methods for creating efficient geroprotectors and biomarkers.

There has been only one known attempt to use AI to predict aging biomarkers, which involved training neural networks on a large age-labeled sample of blood tests [82].

4.2 Aging biomarkers as a computational problem

Aging biomarkers are quantitative characteristics that predict the future life expectancy of an organism based on its current state [72]. They can be normalized to a "biological age," which can be older or younger than the actual age. Future life expectancy is the difference between the average median life expectancy for a species[2] and the biological age of an individual. Different aging biomarkers have different predictive power [64]. For example, gray hair is a marker of aging, but it has low correlation with mortality. Good aging biomarkers should be causally connected to a potential cause of death. Hair color is not causally connected to a potential cause of death, as one could dye one's hair without affecting life expectancy. In contrast, blood pressure and a number of genetic mutations are causally connected with mortality. Thus, they are better biomarkers for aging. Since aging is a complex process, it cannot be expressed by a single number; a large array of parameters is needed to represent it. Aging biomarkers should also be reversible: if the aging process has been reversed, the biomarkers' respective characteristics should change correspondingly (e.g., decrease in number).

There are two ways to find biomarkers: modeling of aging processes, and statistics. As a side note, one could also measure small changes in the Gompertz curve of mortality, that is, use the number of deaths in a population as an aging biomarker [79]. However, to observe them, information about millions of people would be required.

With the help of modern wearables, it is possible to record all the drugs and treatments received by a patient. A huge number of patient records, along with corresponding data on personal genetics, physical movement, and lifetime behavioral activity, could be collected and centralized. This would result in a cohort study with better information supply and stronger probative value. Collecting and interpreting this information would likely require powerful AI.

One plausible AI scenario in biomarker detection is the use of unsupervised machine learning over a large set of biomedical parameters that may lead to the discovery of groups of parameters that correlate with biological aging.

Further, parameter-variance analysis will help to detect real aging biomarkers. For example, the company Gero focuses on gene-stability networks [56].

Another application of AI in the fight against aging is in creating completely new geroprotectors by analyzing cell models, aging models, and molecular properties. Rather than drugs, the geroprotectors could be genetic interventions, that is, insertions of new genes or alterations in the expressions of existing genes (*epigenomics*).

Five hundred thousand British senior citizens have donated their blood and anonymized their healthcare data for use by Biobank, which is now sequencing their genomes. Biobank will provide open access to all the resulting data, which will become an enormous data set for various forms of machine-learning research [125]. Especially promising is the search for genetic networks of aging. Similar projects are taking place in Iceland [81] and Estonia.

4.3 Geroprotector's combinatorial explosion

A number of medications can extend the life of a mouse by slowing down its aging processes [57]. Most of these medications, however, yield only a 10–15% increase in life span. In humans such medications would yield even less, perhaps around 5%, as longer lives are more difficult to extend, and they respond less to known geroprotectors. But what if several geroprotectors are combined? Results of a few studies on mice are promising, as they show a multiplication of effects [96].

Recent research used a sophisticated testing algorithm to identify three drugs that yield maximum life extension in worms and flies [31]. While that algorithm was designed manually, we expect that the best testing scheme would involve AI-aided design of a range of algorithm alternatives.

Although combining certain pairs of geroprotectors works well enough, some geroprotectors are incompatible with one another. Moreover, combining them greatly reduces their effects. Hence, pairwise testing of geroprotector combinations is needed to begin with, followed by larger combinations. To test all combinations of 10 geroprotectors would require 1024 experiments, and for 20 geroprotectors the number of experiments would be over a million, and that is for a single dosage rate for each geroprotector. This is virtually impossible, as there financing has been unsuccessful for even simple testing of one combination on mice (see lifespan.io campaign [126]).

The problem of searching in an enormous space is similar to that of playing a complex board game with a huge search space, such as Go. The recent success of AlphaGo [127] promises that such a search could be simplified. Consequently, a much smaller number of experiments would need to be run to determine an optimal geroprotector combination. The underlying principle of AlphaGo is that the most promising combinations are selected by a neural network trained on a large number of previous games. Similarly, a neural

[2] Technically the life expectancy should be at the biological age, rather than at birth as is usually quoted.

network can be trained to predict the biological effects of chemicals based on knowledge of their properties obtained from a comprehensive library of substances. A similar computational approach is used for drug discovery [92] and toxicity forecasting [103]. *Toxcast* is a large US-government-sponsored program designed to use machine learning to predict the toxicity of different chemicals [86].

To increase the number of useful outcomes of an experiment, it is also necessary to record a vast number of various vital parameter measurements of an organism (for instance, blood composition, physical movement, EEG readings) during the process of geroprotector testing. This would allow the discovery of aging biomarkers during geroprotector testing.

Generally, the geroprotector-identification problem can be reduced to the task of finding a global minimum of a function of ten (or more) variables. A number of efficient machine-learning algorithms are suited for such a task.

The search for aging biomarkers can be pursued in a similar manner. From the mathematical point of view, it is a search for the global minimum of the function of many properties of an organism. The same process can also be used to calculate specific gene interventions for an individual human, in view of the genome characteristics, age, and biomarkers.

Activities in this area are carried out by Gero, Calico, the Buch Institute [19], and others. João Pedro de Magalhães has used random-forest machine learning to predict the properties of life-extending compounds [9].

Additionally, several projects are searching in large combination spaces by using neural networks designed for other tasks:

- Project AtomNet [3] predicts the properties of chemical materials using convolutional neural networks;
- E. Pyzer-Knapp et al. [83] are using a multilayer neural network to predict the electrical properties of new molecules;
- L. Rampasek and A. Goldenberg [84] are reviewing applications of neural-network project TensorFlow by Google in computational biology;
- K. Myint and X.-Q. Xie are predicting ligand properties using a fingerprint-based neural network [74].

4.4 AI, aging, and personalized medicine

Aging can be viewed as the accumulation of errors and lack of adequate regulation in a body by repair mechanisms and the immune system [37]. Hence, in the fight against aging, additional regulation is needed in the form of medical therapy. Medical therapy consists of tests (for instance, blood work, blood pressure readings, medical scans), hypothesizing about causes of disease (diagnosis), medical intervention, and in the case of an incorrect hypothesis, subsequent correction based on new observations.

This process is similar to the scientific method, and at its core it is an information-based process, that is, a process of solving a particular computational task. This means that it will benefit from more data and more intelligent processing, followed by a precise and targeted intervention. Therefore, to cure a disease or rejuvenate a body, it is helpful to collect a large amount of information from that body, in order to construct a detailed model of it. This will enable calculations for the genetic interventions that will lead to recovery and functional improvement.

It is now possible to obtain large amounts of data on a body via full genome sequencing, thousands of parameters of blood analysis, and analysis of the transcriptome, metabolome, and other similar "*omics*" (that is complex quantitative description of functions and statistics of a type of organism's elements). This is achieved through continuous monitoring of food intake, physical activity, and heart parameters via ECG, various scans, and digital tomography. The rapid decline in the cost of all these procedures (US$999 in 2016 for complete sequencing of a genome [78]) has led to individual humans becoming sources of big data. Now we are faced with the question of how to interpret these data to produce the best effects on human health by not only diagnosing existing illnesses but also by predicting future illnesses and creating personalized aging profiles. For this reason, there needs to be better means to derive meaningful conclusions from this vast amount of data.

In the past, the following situation was typical: a patient complains to a doctor about various aches and, after having their blood pressure and temperature measured, receives treatment with a single prescribed medication. In this case the information exchange between the patient and the doctor consisted of "just a few bytes" and some intuitive impressions of the doctor. However, nowadays the information exchange may consist of gigabytes of information at the same cost. For the processing of this data stream, powerful data crunch techniques are required.

During aging, a body gradually accumulates errors, and its natural repair systems begin to fail. The information theory of aging could be designed to enable therapies to correct all these errors, and this idea is at the core of the Strategies for Engineered Negligible Senescence (SENS) project [27].

AI may help humans to model aging by creating a complex causal map of aging processes in a body [90] and then personalizing the model.

Naturally, an organism's body is able to solve most of its problems locally without sending information outside: cells know what to repair, and higher-level attention is needed only when they fail locally. An aging body fails to solve its problems locally. Therefore, it may be reasonable neither to extract information from the body nor to direct therapy into the body, but rather to introduce "AI helpers" inside the body, where they can help solve problems as they appear. Implants and future nanomedicine will be employed along these lines.

Another solution to the "messy problem of aging" is growing completely new body parts and full bodies. However, designing the immunogenic properties of such parts and solving a complex "connection problem" will require analysis of large amounts of information, which will only be feasible if AI is employed.

4.5 Narrow AI in medical-cost reduction and affordable healthcare

Efficient and affordable healthcare will be essential to a global increase in life expectancy. Cheap mobile phones solved the communication problem at the global scale by operating as a standard solution. A similar kind of solution must be sought in healthcare.

High-quality healthcare is very expensive. Nursing, hospitals, drugs, tests, insurance, and highly paid specialists all cost much money, and as a result, advanced healthcare is out of reach for many people.

AI will provide less expensive services and make them available to larger population groups in developing countries. Just as generic drugs can be taken in place of expensive brand-name drugs, an AI-powered consultation could provide diagnostics for people who cannot afford a doctor.

Many people—for instance, those who search the Internet for answers to their medical questions—may be less reluctant to consult an AI-powered specialist than a real doctor.

The following instruments will make AI-based healthcare an inexpensive alternative to hospitals:

- AI chatbots, such as the Babylon app [5];
- Smartphones as a universal diagnostic implement (they can be used to monitor heart rate, diet, physical activity, oxygen saturation, mole changes, and so on);
- Home delivery of cheap generic drugs;
- Web-based medical expert systems.

4.6 Effects of narrow AI on life extension

Narrow AI will help unleash the full potential of life extension, leading to dramatically slower aging. If humans did not age, they could live hundreds of years despite accidents (If we exclude age-dependent component of mortality by extrapolating of minimal probability of death found in 10 years old American girls, which is 0.000084 for a year [1], we will get life expectancy of 5925 years. But increasing probability of death with age lowers it to 81. Most of this death probability increase comes from biological aging.) Yet introduction of narrow AI into effective medical practice could take much longer than related advances in research labs, possibly decades.

The present era of narrow AI might be long, lasting until 2075 by pessimistic predictions [73]. However, this time can be spent usefully, exploring aging biomarkers and geroprotector combinations.

For those who are not able to survive until the arrival of radical life-extension technologies, narrow AI may still play an important role by providing two main backup options: *cryonics* and *digital immortality*.

In cryonics, AI applications may, via wearables, warn a patient's cryonics organization of the impending death of that patient. Cryopreservation could be called plan B, while plan A is to survive until the implantation of life-extension technology.

Digital immortality [107] is the concept of preserving a human being's data so that future AI will be able to reconstruct his or her model using DNA, video recordings, and additional data gleaned from such sources as social networks. It depends on certain assumptions about AI's capabilities, amounts of required information, and the nature of human identity. AI could help to collect and preserve data for digital immortality and perform initial analysis of that data. Digital immortality is plan C in achieving radical life extension.

An early arrival of advanced forms of AI may make these three approaches obsolete before they are implemented.

5 Prospective applications of AGI to life extension

5.1 Personal robot physician

AGI may appear in the form of a human-mind upload [23], [45], or as an infrahuman robotic brain [17] capable of performing most human tasks. It will be Turing complete [112], meaning that it will be able to interact conversationally approximately as well as a human.

There are numerous ways in which AGI may be applied to life extension. In this section, we will explore those that are likely to provide the biggest gains in life expectancy.

Cheap and efficient AGI will enable accessible and predictive personal healthcare. A plausible example is an AI-based personal assistant that will be a combination of a healthcare researcher and personal physician and will be able to provide personal treatment and early response to symptoms. It will constantly monitor an individual's aging biomarkers and other life parameters, allowing daily therapy adjustments. A patient will no longer need to visit a clinic, get a prescription, have it filled at a pharmacy, remember to take drugs at prescribed times, try to determine whether her or she is feeling better, and so on. A personal robot will simply utilize data gathered from wearable monitoring systems to determine an ideal drug combination, order it to be delivered, and then prompt the patient to take a pill. The process of diagnosis and cure will be as effortless and automated as an upgrade of antivirus software on a personal computer.

The ability of AGI to comprehend human language will lead to the possibility of "artificial scientists" that are able to formulate hypotheses, organize experiments, and publish results as scientific papers with less and less help from humans. Combined with robotized labs and less expensive equipment manufacturing, AGI will accelerate scientific research in all fields, including life extension.

Domestic medical robots and wearables will automate clinical trials, reducing costs and accelerating drug discovery by collecting data for clinical trails. Currently, a clinical trial may cost hundreds of millions of dollars because of legal and organizational issues. Home robots will record patient activity, automating clinical trials and making them independent of large medical companies via decentralization, which will reduce their costs and improve data objectivity.

Robotic drones with drugs and defibrillators will provide assistance to people whose wearable systems report an emergency. Domestic robots will monitor the health of a family, help with treatment, monitor medicine consumption, act as physical-exercise instructors, and predict disease. Additionally, they will provide companionship for the elderly, which will also increase life span.

5.2 Integration of monitoring systems into human bodies and nanomedicine

A person's immune system maintains information on such parameters as locations of body inflammation and the types of viruses it is equipped to neutralize. This information is beyond the control of human consciousness. The immune system can be trained with vaccines, but information exchange between humans and immune systems is limited. If a person could read the immune system's information and upload new information into the system, then it would be possible to cure a large range of ailments, including autoimmune diseases, infections, organ failure, tumors, and tissue senescence. Ray Kurzweil expects communication to appear in the 2020s [85]. The process will be similar to current computerized automobile diagnostics. A system of communication between an organism's immune system and a computer can be called a "humoral interface" and would have much in common with a neurointerface. It could be created with some form of nano- or biotechnology, such as computer-programmed cells.

The next step in this direction is *artificial human immune system management*. Such a system may consist of biological organisms, an individual's own upgraded cells [30], or micro robots circulating in an organism's blood. The following are the expected levels of a nanotechnology-based upgrade of the human body:

1) In the first stage, the system will monitor emerging diseases;

2) In the second stage, the system will assist in treatment by killing bacteria, viruses, and cancer cells, and by repairing vascular injuries;

3) In the advanced stages, the system will constantly carry out body repair and treatment of aging;

4) In the final stage, these systems will transform into nanomachines that will replace human cells, making the human body completely artificial and immortal. This will likely only happen when AI reaches the superhuman level.

5.3 "The Upgrade Net": a path to superintelligence through a network of self-improving humans and humanlike AI Systems

As Elon Musk famously tweeted, "Humans must merge with machines or become irrelevant in AI age" [55]. Such a merger would require a powerful **brain-computer interface** (**BCI**), and we think that the best way to achieve this is through the implementation of a personal AI health assistant, which would be integrated into human bodies and brains and focused on preserving human lives.

Musk has also stated [102] that he wants to commercialize the AI health assistant with his Neuralink project. Neuralink will begin by using a simple BCI to treat depression and other mental illnesses. A simple BCI may be used to control human emotions, preventing mental-state-dependent types of violence such as road rage and suicide. This will provide experience that can be directed toward curing mental diseases with BCI, and eventually proceeding to a stage of *augmented humans*, who could later be connected into a network of self-improving humans.

In our opinion, there is another way of building a network of self-improving humans, and it starts with the creation of a medical social network:

First, new type of **patient organizations** [42] will need to be established to connect people who are interested in the fight against aging [128]. These organizations will essentially operate as social networks for information exchange, mutual support, clinical trials, crowdfunding, data collection for digital immortality, civil science, aid in cryopreservation, and political action.

Individual biohackers also could play important role by self experimentation, like Elizabeth Parrish: they could take higher risk experiments on themselves without legal restriction and costs [68].

The next step will be the creation of a network for direct interaction between the brains of human participants, a so-called *neuroweb* [60]. Information-transmission mechanisms may be implemented using weak AI systems. The result of such a network will effectively be a **collective brain**. Direct brain connection may be confusing and inefficient, so a kind of AI firewall may be required to control access to the information that an individual wants to share. Also, an AI dispatcher may be needed to facilitate conversation by remembering conversation's lines, providing relevant links, illustrating ideas, and so on. At a further stage of development, an AGI-based virtual assistant connected through BCI to a human's brain may work as a form of *exocortex* [14].

The ultimate step is to **merge with AI**, which implies blurring the boundaries between the biological brain and the computer. This is equivalent to achieving practical immortality (if no global risks will happen), because brain data will be easily backed up and, if needed, restored. Effectively, human minds and computer superintelligence will merge into a single system. At the same time, people will be able to maintain a preferred level of autonomy with regard to memory, consciousness, and learned skills [34], [101], [75].

6 Superintelligence and the distant future

6.1 Superintelligence finally solving problems of aging and death

We can use trends and polls to predict narrow AI and AGI. Superintelligence is by definition unpredictable. For expectations of its arrival and what it will be able to accomplish, we can refer to various futurists: Bostrom [16], Yamploskiy [113], Yudkowsky [114], Kurzweil [58], Vinge [106], and Goertzel [39] all depict a future dominated by global superintelligence.

According to these futurists, the arrival of superhuman AI will enable solutions to the problems of aging, curing presently incurable diseases, designing universal medical nanorobots, and uploading an individual's consciousness into a computer network.

In the past, it took decades to accomplish complex, globally valuable tasks such as the development of modern aeronautics, wireless communication, and noninvasive surgery; superintelligent AI will be able to solve such problems very quickly, perhaps in moments. With the arrival of superintelligent AI, achieving practical immortality for the majority of people will become feasible.

6.2 Simultaneous creation of superintelligence and advanced nanotechnologies

K. Eric Drexler's book *Engines of Creation* [32] and Robert A. Freitas Jr.'s *Nanomedicine, Volume IIA: Biocompatibility* [36] discuss *nanotechnology* as nanorobotics based on molecular manufacturing for medical treatment and intervention. According to Drexler, medical nanobots will:

- be self-replicating;
- be externally controlled;
- carry onboard computers;
- be capable of swarm behavior
- be cell sized;
- be capable of 3-D printing organic structures; and

be capable of sensing their environment and navigating in it.

If such nanobots arrive before AGI, they will quickly help us map the structure of the human brain and develop technology to create a very powerful supercomputer, leading to the advent of AGI. On the other hand, if AGI arrives first, it will create nanobots. The wait between nanorobotics and AGI will likely be no more than a few years.

Designing the first nanobot and controlling nanorobotic swarms will be a huge computational task, itself requiring the use of available AI.

When this technology matures, it may enable relatively quick (hours to weeks) and seamless replacement of living cells in a human body—with the possible exception of the neurons responsible for personal experiences—with fully controlled nanomachines by injecting a single self-replicating nanobot. Such a nanotechnological body will not age as it will be able constantly self-repair according to original plan.

6.3 Superintelligence and the solution to the consciousness problem: identity copying

On the one hand, it will be difficult to develop full-fledged AGI without first solving the problem of consciousness. On the other hand, nanotechnology and AGI will give us the means to carry out various experiments on the conscious brain and map its structure. For example, investigation of qualia is feasible through a gradual uploading process similar to the thought experiment performed by David Chalmers [22]. This will enable detection of the brain parts and internal processes responsible for subjective experience.

There are two possible scenarios: either there is no mystery here and the problem of uploading consciousness to a computer is purely informational, or consciousness has a certain substrate. This substrate could be a quantum process, continuity of causal relationships, special particles, or similar—that provides identity, and its preservation and transfer is a separate technical task. In either case, the transfer of consciousness to a new carrier is possible: an ordinary computer can be used in the first scenario; the second scenario will require a specialized computer, such as an artificial neuron or a quantum computer[2].

This hypothetical consciousness-receptacle computer will need to be extremely resistant to damage and have advanced backing-up abilities in order to lower the risk of death.

6.4 Using advanced forms of superintelligence for the reconstruction of the dead people

Cryonics is the idea, introduced by Robert Chester Ettinger and Jean Rostand [35] of using low temperatures to preserve human bodies after death until it becomes possible to return them to life. Currently around 250 people are cryopreserved by three cryocompanies [67]. At first, it was thought that bodies could be gradually unfrozen upon the appearance of appropriate technologies. Later it was thought that nanotechnology could be used to repair damage in thawing bodies [32]. A more recent view is that bodies can be scanned without thawing [65]. Advanced tomography [48] or slicing [43] would be employed, and the data from the scans would be entered into a computer, where the human mind would be reconstructed. Currently around 250 people are cryopreserved by three cryocompanies [122] and advanced nanotech created by AI could be used to scan and upload their minds.

In addition, highly evolved superintelligence will be able to reconstruct humans who lived in the past by modeling their lives in a simulation. A reconstruction

would be based on a subject's informational traces. It is called "digital immortality" [108].

For global resurrection of the dead [123], superintelligence may perform a large-scale simulation of the past [124]. Then, based on all the data about the past, it will reconstruct everyone who ever lived.

7 Discussion: strategies for applying AI to life extension

7.1 Problems of AI application in healthcare

In 1979, a rule-based expert system could make a diagnosis better than human doctors [18]. Since then, decades have passed, and yet a large-scale AI revolution still has not happened in healthcare. Most modern medical systems are still based on extremely simple algorithms, for example, *if the heart rate is more than X, execute Y* [8].

Brandon Ballinger [8] wrote that one major obstacle is the majority of "cheap" easily available datasets is not labeled, but machine-learning algorithms mostly require labeled data for training. For example, there is a lot of cardiac data, but it is not clear what disease it is associated with or what the patient's vital parameters were. To obtain labeled data, it might be necessary to conduct costly and potentially harmful experiments on humans. Currently, this problem is being approached by unsupervised learning algorithms, which do not require labeled data, but their performance is still behind that of the supervised systems.

In addition, there are regulatory issues regarding the utilization of AI in healthcare, as well as disputes about risk allocation and insurance payments between startups and hospitals. AI can easily be migrated into an individual's smartphone, but getting it into a doctor's office is more complicated, not to mention the intricacies of accounting for AI in insurance payment systems.

One can imagine that the modest pace of advancement of AI applications in healthcare in recent decades might be disappointing to the authors of the first edition of *Artificial Intelligence in Medicine*, which was published back in 1982 [97]. Yet, due to substantial increase in computing power, availability of "cheap" digitized data, advanced data-analysis algorithms, and new regulations, we finally seem to find ourselves at the dawn of the rapid development of AI in healthcare.

Privacy issues regarding personal data create a trade-off for AI development. On one hand, the greater the amount of open data, the easier it is to train AI algorithms. (Sharing one's personal health data may cause unpredictable harm to the individual, however.) On the other hand, if only anonymized data is available, important vital parameters and data points will be lost. The patient organizations discussed in section 5.3 may understand the importance of providing open access to personal data, as doing so would help train AI for healthcare.

7.2 AI in medicine, and AI safety

Issues of AI safety, on both local and global levels, are beyond the scope of this work. We want to emphasize just two points of intersection of AI in healthcare and AI safety:

Medical AI is aimed at the preservation of human lives, whereas, for example, military AI is generally focused on human destruction. If we assume that AI preserves the values of its creators, medical AI should be more harmless.

The development of such types of medical AI as neuroimplants will accelerate the development of AI in the form of a distributed social network consisting of self-upgrading people. Here, again, the values of such an intelligent neuroweb will be defined by the values of its participant "nodes," which should be relatively safer than other routes to AI. Also, AI based on human uploads may be less probable to go into quick unlimited self-improvement, because of complex and opaque structure.

If the orthogonality of values and intelligence thesis [16] has some exceptions, medical AI may be safer than military AI.

On the other way, medical AI may increase the risks as it will open the way to the neuromorphic AI, which is regarded dangerous [16], or it will be under less control than military AI, and could run into explosive run-away self-improvement.

The Upgrade Net discussed above may become a useful instrument in solving the AI safety problem, as the growing collective human intelligence could operate as a global police force, identifying potential terrorist behavior and other threats.

The safety will come from intrinsic value alignment of human uploads [94], combined with superintelligence power of the whole net which will be able to find and prevent appearance of other types of potentially dangerous AI systems, as well as exterminate the need of creation of such systems. Turchin addressed this question in greater details in [99].

7.3 Surviving to see AGI: personalized, age-dependent strategies

The older a person gets, the lower his or her chances of surviving into the era of AGI and powerful life-extension technologies. Fortunately, it is not necessary to wait until superintelligence arises. In order for an individual's life expectancy to be increased indefinitely, that individual must stay alive only until the moment when average life expectancy begins increasing by more than a year each year, at which point longevity escape velocity will be achieved [27].

However, the chances that a person will be able to benefit from life extension significantly increase if that person has better access to upcoming technologies by, for instance, living in a developed country, having financial security, or being foresighted enough to research and utilize those technologies when first available.

In order to increase and spread the benefits of medical AI in the future, it will be necessary to increase

people's awareness and encourage them to exercise all available means for life extension. As part of this strategy, we promote participation in patient organizations committed to fighting aging, signing up for cryonics, and sharing and collecting digital immortality data.

8 Conclusion

This work is an overview of the existing and prospective AI applications that the authors consider the most promising and beneficial for life extension and antiaging. We have considered a wide range of problems with the current state of the research and the industry, the most promising prospective applications of AI, and strategies to increase public awareness in order to ensure maximal life-extension opportunities for everyone.

Based on related work, we have reviewed the expected stages of the development of AI in the near future, and estimated when the most advanced levels will arrive.

Further, we have presented an overview of the current AI-based healthcare projects of certain for-profit companies of various scales. These projects include IBM Watson Healthcare, Google Calico, and DeepMind Health, as well as the research projects of certain academic groups and nonprofit organizations.

We have shown that the exponential growth of AI's capabilities makes it more likely that AI could help fight the exponential increase of the probability of a human being's mortality over time, and that AI could help a person to reach longevity escape velocity before superintelligence is achieved. It may help millions or maybe even billions of people to "survive until immortality," and thus rescue their life from impending death. Some of the authors explored this topic in greater detail in the article "Fighting aging as an effective altruism case: the model of impact" [100].

We have emphasized the importance of establishing patient organizations to spread awareness of the subjects of life extension, voluntary patient data collection, early adoption of medical AI technologies, and the eventual formation of a "neuroweb" with the arrival of advanced forms of AI.

9 Acknowledgements

We thank Anastasia Egorova, Maxim Cholin, Sergei Shegurin, Alexandra Alexeeva, and Dmitriy Shakhov for interesting discussions that helped in the preparation of the article.

Conflict of interest: No.

Research funding: No.

10 References

[1] Actuarial Life Table. 2017. Actuarial Life Table. Retrieved August 24, 2017 from https://www.ssa.gov/oact/STATS/table4c6.html

[2] Victor Yu Argonov. 2012. Neural Correlate of Consciousness in a Single Electron: Radical Answer to "Quantum Theories of Consciousness." NeuroQuantology 10, 2 (2012). Retrieved from http://neuroquantology.com/index.php/journal/article/view/548

[3] Atomnet. 2017. Introducing AtomNet – Drug design with convolutional neural networks. Retrieved from http://www.atomwise.com/introducing-atomnet/

[4] DN Aubrey. 2004. Escape velocity: why the prospect of extreme human life extension matters now. PLoS Biol. 2, 6 (2004), e187.

[5] Babylon Health. 2017. Retrieved from https://www.babylonhealth.com/

[6] Xiao-chen Bai, Thomas G Martin, Sjors HW Scheres, and Hendrik Dietz. 2012. Cryo-EM structure of a 3D DNA-origami object. Proc. Natl. Acad. Sci. 109, 49 (2012), 20012–20017.

[7] Katherine Bailey. 2017. Conversational AI and the road ahead. Retrieved from https://techcrunch.com/2017/02/25/conversational-ai-and-the-road-ahead/

[8] Brandon Ballinger. 2016. Three Challenges for Artificial Intelligence in Medicine. Retrieved from https://blog.cardiogr.am/three-challenges-for-artificial-intelligence-in-medicine-dfb9993ae750

[9] Diogo G. Barardo, Danielle Newby, Daniel Thornton, Taravat Ghafourian, João Pedro de Magalhães, and Alex A. Freitas. 2017. Machine learning for predicting lifespan-extending chemical compounds. Aging (2017).

[10] David J Barnes and Dominique Chu. 2010. Introduction to modeling for biosciences. Springer Science & Business Media.

[11] K. Belcher. 2016. From $600 M to $6 Billion, Artificial Intelligence Systems Poised for Dramatic Market Expansion in Healthcare. Frost & Sullivan. Retrieved August 24, 2017 from https://ww2.frost.com/news/press-releases/600-m-6-billion-artificial-intelligence-systems-poised-dramatic-market-expansion-healthcare/

[12] Jamie Bennett, Osvaldas Rokas, and Liming Chen. 2017. Healthcare in the Smart Home: A Study of Past, Present and Future. Sustainability 9, 5 (2017), 840.

[13] Kent S Boles, Krishna Kannan, John Gill, Martina Felderman, Heather Gouvis, Bolyn Hubby, Kurt I Kamrud, J Craig Venter, and Daniel G Gibson. 2017. Digital-to-biological converter for on-demand production of biologics. Nat. Biotechnol. 35 672–675 2017 (2017).

[14] Tamara Bonaci, Jeffrey Herron, Charlie Matlack, and Howard Jay Chizeck. 2014. Securing the exocortex: A twenty-first century cybernetics challenge. 1–8.

[15] N. Bostrom. 2003. Astronomical waste: The opportunity cost of delayed technological development. Utilitas 15, 3 (2003), 308–314.

[16] N. Bostrom. 2014. Superintelligence. Oxford University Press, Oxford.

[17] N. Bostrom and A. Sandberg. 2008. Whole brain emulation: a roadmap. Lanc. Univ. Accessed January 21, (2008), 2015.

[18] Bruce G Buchanan and Edward Hance Shortliffe. 1984. Rule-based expert systems. Addison-Wesley Reading, MA.

[19] Buck Institute. Home | The Buck Institute for Research on Aging. Retrieved August 24, 2017 from https://www.buckinstitute.org/

[20] Filipe Cabreiro, Catherine Au, Kit-Yi Leung, Nuria Vergara-Irigaray, Helena M Cochemé, Tahereh Noori, David Weinkove, Eugene Schuster, Nicholas DE Greene, and David Gems. 2013. Metformin retards aging in C. elegans by altering microbial folate and methionine metabolism. Cell 153, 1 (2013), 228–239.

[21] Hugo Campos. 2009. Soccer player Anthony Van Loo survives a sudden cardiac arrest (SCA) when his ICD fires. Retrieved from https://www.youtube.com/watch?v=DU_i0ZzIV5U

[22] D. Chalmers. 1996. The Conscious Mind. Oxford University Press, New York.

[23] William P Cheshire JR. 2015. The Sum of All Thoughts: Prospects of Uploading the Mind to a Computer. Ethics Med. 31, 3 (2015), 135.

[24] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Wei Xie, and Gail L Rosen. 2017. Opportunities And Obstacles For Deep Learning In Biology And Medicine. bioRxiv (2017). Retrieved from http://www.biorxiv.org/content/early/2017/05/28/142760

[25] M Coppock. 2017. Researchers Are Using Neural Networks To Get Better At Reading Our MindsR MINDS. Retrieved from https://www.digitaltrends.com/computing/researchers-use-neural-network-algorithms-for-more-accurate-brain-scans/

[26] Brad Darrow. 2016. Why IBM Is Dropping $2.6 Billion on Truven Health. Retrieved from http://fortune.com/2016/02/18/ibm-truven-health-acquisition/

[27] Aubrey De Grey and Michael Rae. 2007. Ending aging: The rejuvenation breakthroughs that could reverse human aging in our lifetime. St. Martin's Press.

[28] Stephan De Spiegeleire, Matthijs Maas, and Tim Sweijs. 2017. Artificial intelligence and the future of defence. Retrieved from http://www.hcss.nl/sites/default/files/files/reports/Artificial%20Intelligence%20and%20the%20Future%20of%20Defense.pdf

[29] Deepmind. 2017. Helping clinicians get patients from test to treatment, faster. Retrieved from https://deepmind.com/applied/deepmind-health/

[30] Linda Delacey. 2017. Cyborg step? Scientists engineer bioelectric cells. New Atlas. Retrieved from http://newatlas.com/cyborg-technology-scientists-create-bioelectric-hybrid-cells/47481/

[31] Tesfahun Dessale, Krishna Chaithanya Batchu, Diogo Barardo, Li Fang Ng, Vanessa Yuk Man Lam, Markus R Wenk, Nicholas S Tolwinski, and Jan Gruber. 2017. Slowing ageing using drug synergy in C. elegans. bioRxiv (2017), 153205.

[32] Drexler. 1986. E.: Engines of Creation. Anchor Press.

[33] [33] Petter Eckersley and Nasser Yomna. 2017. Measuring the progress of AI research. Retrieved from https://www.eff.org/ai/metrics

[34] Douglas C Engelbart. 1962. Augmenting human intellect: a conceptual framework (1962). Pack. Randall JORDAN Ken Multimed. Wagner Virtual Real. N. Y. WW Nort. Co. (1962), 64–90.

[35] Robert CW Ettinger and Jean Rostand. 1965. The prospect of immortality. Sidgwick and Jackson.

[36] Robert A Freitas Jr. 2003. Nanomedicine, Vol. IIA: Biocompatibility. Landes Biosci. Georget. USA (2003).

[37] Leonid A Gavrilov and Natalia S Gavrilova. 2001. The reliability theory of aging and longevity. J. Theor. Biol. 213, 4 (2001), 527–545.

[38] Carmine Giardino, Xiaofeng Wang, and Pekka Abrahamsson. 2014. Why early-stage software startups fail: a behavioral framework. 27–41.

[39] Goertzel. Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood? J. Conscious. Stud. 19 No 1–2 2012 Pp 96–111. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.352.3966&rep=rep1&type=pdf

[40] Yelena Gorina, Donna Hoyert, Harold Lentzner, and Margie Goulding. 2006. Trends in Causes of Death among Older Persons in the United States. Retrieved from https://www.cdc.gov/nchs/data/ahcd/agingtrends/06olderpersons.pdf

[41] K. Grace. 2017. When Will AI Exceed Human Performance? Evidence from AI Experts. Retrieved from https://arxiv.org/pdf/1705.08807.pdf

[42] N Guillamón, M Armayones, E Hernández, and B Gómez-Zúñiga. 2010. The role of patient organizations in participatory medicine: Can virtual health communities help participatory medicine accomplish its objectives. J. Particip. Med. 2, (2010), e21.

[43] Gwern. 2017. Plastination versus Cryonics. (2017). Retrieved from https://www.gwern.net/plastination

[44] Nichois Hannah. 2017. The top 10 leading causes of death in the United States. Retrieved from http://www.medicalnewstoday.com/articles/282929.php

[45] R Hanson. 2016. The Age of Em: Work, Love, and Life when Robots Rule the Earth. Oxford University Press.

[46] Erika Check Hayden. 2016. A $3-billion plan to cure disease. (2016).

[47] Erika Check Hayden. 2016. Biology software promises easier way to program living cells. Nat. News (2016). Retrieved from Biology software promises easier way to program living cells

[48] Kenneth J Hayworth. 2012. Electron imaging technology for whole brain neural circuit mapping. Int. J. Mach. Conscious. 4, 1 (2012), 87–108.

[49] Clyde A Hutchison, Ray-Yuan Chuang, Vladimir N Noskov, Nacyra Assad-Garcia, Thomas J Deerinck, Mark H Ellisman, John Gill, Krishna Kannan, Bogumil J Karas, and Li Ma. 2016. Design and synthesis of a minimal bacterial genome. Science 351, 6280 (2016), aad6253.

[50] IBM. 2017. Medical Sieve. Retrieved from http://researcher.watson.ibm.com/researcher/view_group.php?id=4384

[51] CB Insights. 2016. From Virtual Nurses to Drug Discovery: 65+ Artificial Intelligence Startups in Healthcare. CB Insights (2016).

[52] Intel. 2017. Intel healthcare overview. Retrieved from http://www.intel.com/content/www/us/en/healthcare-it/healthcare-overview.html

[53] John PA Ioannidis. 2005. Contradicted and initially stronger effects in highly cited clinical research. Jama 294, 2 (2005), 218–228.

[54] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I Glass, and Markus W Covert. 2012. A whole-cell computational model predicts phenotype from genotype. Cell 150, 2 (2012), 389–401.

[55] Arjun Kharpal. 2017. Elon Musk: Humans must merge with machines or become irrelevant in AI age. CNBC. Retrieved from http://www.cnbc.com/2017/02/13/elon-musk-humans-merge-machines-cyborg-artificial-intelligence-robots.html

[56] Valeria Kogan, Ivan Molodtsov, Leonid I Menshikov, Robert J Shmookler Reis, and Peter Fedichev. 2015. Stability analysis of a model gene network links aging, stress resistance, and negligible senescence. Sci. Rep. 5, (2015), 13589.

[57] Maria Konovalenko. 2016. Longevity Cookbook: Combinations of Life Extending Drugs. Retrieved from https://medium.com/@mariakonovalenko/longevity-cookbook-combinations-of-life-extending-drugs-d092feb64c46

[58] Ray Kurzweil. 2006. Singularity is Near. Viking.

[59] Christian de Looper. 2016. Google's smart bathroom patent puts sensors in your toilet, tub, and mirror. Digital trends. Retrieved July 16, 2017 from https://www.digitaltrends.com/home/google-smart-bathroom-patent/

[60] Pavel Luksha. 2014. NeuroWeb Roadmap: Results of Foresight & Call for Action. Retrieved from https://www.slideshare.net/PavelLuksha/neuroweb-roadmap-preliminary

[61] John Markoff. 2011. Computer Wins on "Jeopardy!" Retrieved from http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html?pagewanted=all

[62] Albert W Marshall and Ingram Olkin. 2015. A bivariate Gompertz–Makeham life distribution. J. Multivar. Anal. 139, (2015), 219–226.

[63] Gianluka Mauro. 2016. Six graphs to understand the state of Artificial Intelligence academic research. Retrieved from https://blog.ai-academy.com/six-graphs-to-understand-the-state-of-ai-academic-research-3a79cac4c9c2

[64] Steve Hill May. 2017. The Need for Better Aging Biomarkers. Life Ext. Advocac Found. (2017). Retrieved from http://www.leafscience.org/aging-biomarkers/

[65] Ralf C Merkle. 1994. The molecular repair of the brain. Cryonics Mag. 15, (1994).

[66] Richard A Miller, David E Harrison, Clinton M Astle, Elizabeth Fernandez, Kevin Flurkey, Melissa Han, Martin A Javors, Xinna Li, Nancy L Nadon, and James F Nelson. 2014. Rapamycin-mediated lifespan increase in mice is dose and sex dependent and metabolically distinct from dietary restriction. Aging Cell 13, 3 (2014), 468–477.

[67] Ole Martin Moen. 2015. The case for cryonics. J. Med. Ethics (2015), medethics-2015.

[68] Dara Mohammadi and Nicola Davis. 2016. Can this woman cure ageing with gene therapy? The Observer. Retrieved November 14, 2017 from http://www.theguardian.com/science/2016/jul/24/elizabeth-parrish-gene-therapy-ageing

[69] Molecula Maxima. 2017. Retrieved from https://moleculamaxima.com/

[70] Megan Molteni. 2017. Blockchain Could Be the Answer to Health Care's Electronic Record Woes. Wired. Retrieved July 16, 2017 from https://www.wired.com/2017/02/moving-patient-data-messy-blockchain-help/

[71] A. Moskalev, Elizaveta Chernyagina, Vasily Tsvetkov, Alexander Fedintsev, Mikhail Shaposhnikov, Vyacheslav Krut'ko, Alex Zhavoronkov, and Brian K. Kennedy. 2016. Developing criteria for evaluation of geroprotectors as a key stage toward translation to the clinic. Aging Cell 15, 3 (June 2016), 407–415. DOI:https://doi.org/10.1111/acel.12463

[72] AA Moskalev and MA Batin. 2011. Biomarkers of aging and aging-related pathologies. Dep. Bioeng. Bioinforma. MV Lomonosov Mosc. State Univ. (2011), 63.

[73] Vincent C Müller and N. Bostrom. 2016. Future progress in artificial intelligence: A survey of expert opinion. In Fundamental issues of artificial intelligence. Springer, 553–570.

[74] Kyaw Z Myint and Xiang-Qun Xie. 2015. Ligand Biological Activity Predictions Using Fingerprint-Based Artificial Neural Networks (FANN-QSAR). Artif. Neural Netw. (2015), 149–164.

[75] Miguel AL Nicolelis. 2014. Brain-to-Brain Interfaces: When Reality Meets Science Fiction.

[76] Joshua Ogawa. 2017. Global AI startup financing hit $5bn in 2016. Retrieved from http://asia.nikkei.com/Business/Trends/Global-AI-startup-financing-hit-5bn-in-2016

[77] Open worm. 2017. Retrieved from http://www.openworm.org/

[78] Alexandra Ossola. 2015. Your Full Genome Can Be Sequenced and Analyzed For Just $1,000. Pop. Sci. (2015). Retrieved from http://www.popsci.com/cost-full-genome-sequencing-drops-to-1000

[79] L. Piantanelli, G. Rossolini, A. Basso, A. Piantanelli, M. Malavolta, and A. Zaia. 2001. Use of mathematical models of survivorship in the study of biomarkers of aging: the role of heterogeneity. Mech. Ageing Dev. 122, 13 (September 2001), 1461–1475. DOI:https://doi.org/10.1016/S0047-6374(01)00271-8

[80] Alessandro Polini, Ljupcho Prodanov, Nupura S Bhise, Vijayan Manoharan, Mehmet R Dokmeci, and Ali Khademhosseini. 2014. Organs-on-a-chip: a new tool for drug discovery. Expert Opin. Drug Discov. 9, 4 (2014), 335–352.

[81] Alison Proffitt. 2013. NextCODE Health Launches deCODE's Clinical Genomics Platform. Retrieved July 16, 2017 from http://www.bio-itworld.com

[82] Evgeny Putin, Polina Mamoshina, Alexander Aliper, Mikhail Korzinkin, Alexey Moskalev, Alexey Kolosov, Alexander Ostrovskiy, Charles Cantor, Jan Vijg, and Alex Zhavoronkov. 2016. Deep biomarkers of human aging: application of deep neural networks to biomarker development. Aging 8, 5 (2016), 1021.

[83] Edward O Pyzer-Knapp, Kewei Li, and Alan Aspuru-Guzik. 2015. Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery. Adv. Funct. Mater. 25, 41 (2015), 6495–6502.

[84] Ladislav Rampasek and Anna Goldenberg. 2016. Tensorflow: Biology's gateway to deep learning? Cell Syst. 2, 1 (2016), 12–14.

[85] Lidia Ramsey. 2016. Futurist Ray Kurzweil wants to use tiny robots in our bloodstream to fight disease and live forever. Business insider. Retrieved from http://www.businessinsider.com/ray-kurzweil-on-nanobots-and-the-immune-system-2016-4

[86] Ann M Richard, Richard S Judson, Keith A Houck, Christopher M Grulke, Patra Volarath, Inthirany Thillainadarajah, Chihae Yang, James Rathman, Matthew T Martin, and John F Wambaugh. 2016. ToxCast chemical landscape: paving the road to 21st century toxicology. Chem. Res. Toxicol. 29, 8 (2016), 1225–1251.

[87] Arlan Richardson, Steven N Austad, Yuji Ikeno, Archana Unnikrishnan, and Roger J McCarter. 2016. Significant life extension by ten percent dietary restriction. Ann. N. Y. Acad. Sci. 1363, 1 (2016), 11–17.

[88] A Rosenberg and JH Meyerle. 2017. Total-body photography in skin cancer screening: the clinical utility of standardized imaging. Cutis 99, 5 (2017), 312.

[89] J. Russell. 2017. After beating the world's elite Go players, Google's AlphaGo AI is retiring. TechCrunch. Retrieved from https://techcrunch.com/2017/05/27/googles-alphago-ai-is-retiring/

[90] Andrew D Rutenberg, Arnold B Mitnitski, Spencer Farrell, and Kenneth Rockwood. 2017. Unifying ageing and frailty through complex dynamical networks. ArXiv Prepr. ArXiv170606434 (2017).

[91] Aaron Saenz. 2009. Smart Toilets: Doctors in Your Bathroom. Singularity Hub. Retrieved July 16, 2017 from https://singularityhub.com/2009/05/12/smart-toilets-doctors-in-your-bathroom/

[92] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. 2017. Generating focussed molecule libraries for drug discovery with recurrent neural networks. ArXiv Prepr. ArXiv170101329 (2017).

[93] V. Shakirov. 2016. Review of state-of-the-arts in artificial intelligence with application to AI safety problem. ArXiv Prepr. ArXiv160504232 (2016). Retrieved from https://arxiv.org/abs/1605.04232

[94] Carl Shulman. 2010. Whole brain emulation and the evolution of superorganisms. Mach. Intell. Res. Inst. Work. Pap. Httpintelligence OrgfilesWBE-Superorgs Pdf (2010).

[95] Andrew Sparkes, Wayne Aubrey, Emma Byrne, Amanda Clare, Muhammed N Khan, Maria Liakata, Magdalena Markham, Jem Rowland, Larisa N Soldatova, and Kenneth E Whelan. 2010. Towards Robot Scientists for autonomous scientific discovery. Autom. Exp. 2, 1 (2010), 1.

[96] Randy Strong, Richard A Miller, Clinton M Astle, Robert A Floyd, Kevin Flurkey, Kenneth L Hensley, Martin A Javors, Christiaan Leeuwenburgh, James F Nelson, and Ennio Ongini. 2008. Nordihydroguaiaretic acid and aspirin increase lifespan of genetically heterogeneous male mice. Aging Cell 7, 5 (2008), 641–650.

[97] Peter Szolovits. 1982. Artificial intelligence in medicine. Westview Press Boulder, CO.

[98] Turbine AI. 2017. Retrieved from Turbine.ai

[99] A. Turchin and D. Denkenberger. 2017. Global Solutions of the AI Safety Problem.

[100] A. Turchin, D. Denkenberger, E. Milova, A. Egorova, and M. Batin. 2017. Fighting aging as an effective altruism case: the model of impact.

[101] Valentin Turchin and Cliff Joslyn. 1990. Communications: The Cybernetic Manifesto (Part I). Kybernetes 19, 2 (1990), 63–64.

[102] Tim Urban. 2017. Neuralink and the Brain's Magical Future. Retrieved from http://waitbutwhy.com/2017/04/neuralink.html

[103] ORD US EPA. 2015. Toxicity Forecasting. US EPA. Retrieved July 16, 2017 from https://www.epa.gov/chemical-research/toxicity-forecasting

[104] Monica Van Such, Robert Lohr, Thomas Beckman, and James M Naessens. 2017. Extent of diagnostic agreement among medical referrals. J. Eval. Clin. Pract. (2017).

[105] Jan Vijg and Aubrey DNJ De Grey. 2014. Innovating aging: promises and pitfalls on the road

to life extension. Gerontology 60, 4 (2014), 373–380.

[106] Vernor Vinge. 1993. Technological singularity. 30–31.

[107] Gian Volpicelli. 2016. This Transhumanist Records Everything Around Him So His Mind Will Live Forever. Vice.Motherboard. Retrieved from https://motherboard.vice.com/en_us/article/4xangw/this-transhumanist-records-everything-around-him-so-his-mind-will-live-forever

[108] Gian Volpicelli. 2016. This Transhumanist Records Everything Around Him So His Mind Will Live Forever. Motherboard. Retrieved August 24, 2017 from https://motherboard.vice.com/en_us/article/4xangw/this-transhumanist-records-everything-around-him-so-his-mind-will-live-forever

[109] Brian Wang. 2017. Final Frontier Medical Devices won the tricorder xprize with device that can diagnose 90% of ER situations at home. Retrieved from https://www.nextbigfuture.com/2017/04/final-frontier-medical-devices-won-the-tricorder-xprize.html

[110] John G White, Eileen Southgate, J Nichol Thomson, and Sydney Brenner. 1986. The structure of the nervous system of the nematode Caenorhabditis elegans: the mind of a worm. Phil Trans R Soc Lond 314, (1986), 1–340.

[111] Zongli Xu and Jack A Taylor. 2013. Genome-wide age-related DNA methylation changes in blood and other tissues relate to histone modification, expression and cancer. Carcinogenesis 35, 2 (2013), 356–364.

[112] R. Yampolskiy. 2013. Turing test as a defining feature of AI-completeness. Artif. Intell. Evol. Comput. Metaheuristics (2013), 3–17.

[113] R. Yampolskiy. 2015. Artificial Superintelligence: a Futuristic Approach. CRC Press.

[114] E. Yudkowsky. 2008. Artificial Intelligence as a Positive and Negative Factor in Global Risk, in Global Catastrophic Risks. Oxford University Press: Oxford, UK.

[115] Alex Zhavoronkov and Bhupinder Bhullar. 2015. Classifying aging as a disease in the context of ICD-11. Front. Genet. 6, (2015).

[116] 2013. Intelligent Toilet Monitors Your Health. iStep Blog by ifm efector. Retrieved July 16, 2017 from http://istep.ifmefector.com/2013/04/09/intelligent-toilet-monitors-your-health/

[117] 2016. JCVI: First Minimal Synthetic Bacterial Cell Designed and Constructed by Scientists at Venter Institute and Synthetic Genomics, Inc. Retrieved July 16, 2017 from http://www.jcvi.org/cms/press/press-releases/full-text/article/first-minimal-synthetic-bacterial-cell-designed-and-constructed-by-scientists-at-venter-institute-an

[118] 2017. Global Artificial Intelligence in Medicine Market. Retrieved from https://www.mordorintelligence.com/industry-reports/artificial-intelligence-in-medicine-market

[119] 2017. Deep genomics. Retrieved from https://www.deepgenomics.com/

[120] 2017. Atomwise. Retrieved from http://www.atomwise.com/

[121] 2017. Biohub cell atlas. Retrieved from https://czbiohub.org/projects/cell-atlas/

[122] 2017. Biostamp. Retrieved from https://www.mc10inc.com/our-products/biostamprc

[123] 2017. Nimb: A Smart Ring That Helps You Feel Safe And Sound. Retrieved from https://www.kickstarter.com/projects/1629204423/nimb-a-smart-ring-that-keeps-you-safe-and-sound

[124] 2017. Artificial Intelligence for Continuous Patient Monitoring. Retrieved from https://aicure.com/

[125] 2017. GSK and Regeneron to mine gene data from 500,000 Britons. Reuters. Retrieved July 16, 2017 from http://uk.reuters.com/article/uk-health-genes-gsk-regeneron-pharms-idUKKBN16U01O

[126] 2017. Lifespan.io. Retrieved July 16, 2017 from https://www.lifespan.io/

[127] 2017. AlphaGo. Retrieved from https://deepmind.com/research/alphago/

[128] 2017. Open Longevity. Retrieved from http://openlongevity.org/

Modeling and Interpreting Expert Disagreement About Artificial Superintelligence

Seth D. Baum, Anthony M. Barrett, and Roman V. Yampolskiy
Global Catastrophic Risk Institute, PO Box 40364, Washington, DC 20016, USA
http://gcrinstitute.org,
E-mail: seth@gcrinstitute.org

Keywords: artificial superintelligence, expert judgment, risk analysis

Received: Avgust 31, 2017

Artificial superintelligence (ASI) is artificial intelligence (AI) with capabilities that are significantly greater than human capabilities across a wide range of domains. A hallmark of the ASI issue is disagreement among experts. This paper demonstrates and discusses methodological options for modeling and interpreting expert disagreement about the risk of ASI catastrophe. Using a new model called ASI-PATH, the paper models a well-documented recent disagreement between Nick Bostrom and Ben Goertzel, two distinguished ASI experts. Three points of disagreement are considered: (1) the potential for humans to evaluate the values held by an AI, (2) the potential for humans to create an AI with values that humans would consider desirable, and (3) the potential for an AI to create for itself values that humans would consider desirable. An initial quantitative analysis shows that accounting for variation in expert judgment can have a large effect on estimates of the risk of ASI catastrophe. The risk estimates can in turn inform ASI risk management strategies, which the paper demonstrates via an analysis of the strategy of AI confinement. The paper find the optimal strength of AI confinement to depend on the balance of risk parameters (1) and (2).

Povzetek: Predstavljena je metoda za modeliranje in interpretiranje razlik v mnenjih ekspertov o superinteligenci.

1 Introduction

Artificial superintelligence (ASI) is artificial intelligence (AI) with capabilities that are significantly greater than human capabilities across a wide range of domains. If developed, ASI could have impacts that are highly beneficial or catastrophically harmful, depending on its design

A hallmark of the ASI issue is disagreement among experts. Experts disagree on if ASI will be built, when it would be built, what designs it would use, and what its likely impacts would be.[1] The extent of expert disagreement speaks to the opacity of the underlying ASI issue and the general difficulty of forecasting future technologies. This stands in contrast with other major global issues, such as climate change, for which there is extensive expert agreement on the basic parameters of the issue (Oreskes 2004). Expert consensus does not guarantee that the issue will be addressed—the ongoing struggle to address climate change attests to this—but it does offer direction for decision making.

In the absence of expert agreement, those seeking to gain an understanding of the issue must decide what to believe given the existence of the disagreement. In some cases, it may be possible to look at the nature of the disagreement and pick sides; this occurs if other sides clearly have flawed arguments that are not worth giving any credence to. However, in many cases, multiple sides of a disagreement make plausible arguments; in these cases, the thoughtful observer may wish to form a belief that in some way considers the divergent expert opinions.

This paper demonstrates and discusses methodological options for modeling and interpreting expert disagreement about the risk of ASI catastrophe. The paper accomplishes this by using a new ASI risk model called ASI-PATH (Barrett and Baum 2017a; 2017b). Expert disagreement can be modeled as differing estimates of parameters in the risk model. Given a set of differing expert parameter estimates, aggregate risk estimates can be made using weighting functions. Modeling expert disagreement within the context of a risk model is a method that has been used widely across a range of other contexts; to our knowledge this paper marks the first application of this method to ASI.

The paper uses a well-documented recent disagreement between Nick Bostrom and Ben Goertzel as an illustrative example—an example that is also worthy of study in its own right. Bostrom and Goertzel are both longstanding thought leaders about ASI, with lengthy research track records and a shared concern with the societal impacts of ASI. However, in recent publications, Goertzel (2015; 2016) expresses significant

[1] On expert opinion of ASI, see Baum et al. (2011), Armstrong and Sotala (2012), Armstrong et al. (2014), and Müller and Bostrom (2014).

disagreement with core arguments made by Bostrom (2014). The Bostrom-Goertzel disagreement is notable because both of them are experts whose arguments about ASI can be expected to merit significant credence from the perspective of an outside observer. Therefore, their disagreement offers a simple but important case study for demonstrating the methodology of modeling and interpreting expert disagreement about ASI.

The paper begins by summarizing the terms of the Bostrom-Goertzel disagreement. The paper then introduces the ASI-PATH model and shows how the Bostrom-Goertzel disagreement can be expressed in terms of ASI-PATH model parameters. The paper then presents model parameter estimates based on the Bostrom-Goertzel disagreement. The parameter estimates are not rigorously justified and instead are intended mainly for illustration and discussion purposes. Finally, the paper applies the risk modeling to a practical problem, that of AI confinement.

2 The Bostrom-Goertzel disagreement

Goertzel (2015; 2016) presents several disagreements with Bostrom (2014). This section focuses on three disagreements of direct relevance to ASI risk.

2.1 Human evaluation of AI values

One disagreement is on the potential for humans to evaluate the values that an AI has. Humans would want to diagnose an AI's values to ensure that they are something that humans consider desirable (henceforth "human-desirable"). If humans find an AI to have human-undesirable values, they can reprogram the AI or shut it down. As an AI gains in intelligence and power, it will become more capable of realizing its values, thus making it more important that its values are human-desirable. A core point of disagreement concerns the prospects for evaluating the values of AI that have significant but still subhuman intelligence levels. Bostrom indicates relatively low prospects for success at this evaluation, whereas Goertzel indicates relatively high prospects for success.

Bostrom (2014, p.116-119) posits that once an AI reaches a certain point of intelligence, it might adopt an adversarial approach. Bostrom dubs this point the "treacherous turn":

> *The treacherous turn*: While weak, an AI behaves cooperatively (increasingly so, as it gets smarter). When the AI gets sufficiently strong–without warning or provocation–it strikes, forms a singleton [i.e., takes over the world], and begins directly to optimize the world according to the criteria implied by its final values. (Bostrom 2014, p.119)

Such an AI would not have durable values in the sense that it would go from acting in human-desirable ways to acting in human-undesirable ways. A key detail of the treacherous turn theory is that the AI has values that are similar to, but ultimately different from, human-desirable values. As the AI gains intelligence, it goes through a series of stages:

1. At low levels of intelligence, the AI acts in ways that humans consider desirable. At this stage, the differences between the AI's values and human values are not important because the AI can only complete simple tasks that are human-desirable.
2. At an intermediate level of intelligence, the AI realizes that its values differ from human-desirable values *and* that it if it tried deviating from human-desirable values, humans would reprogram the AI or shut it down. Furthermore, the AI discovers that it can successfully pretend to have human-desirable values until it is more intelligent.
3. At a high level of intelligence, the AI takes control of the world from humanity so that humans cannot reprogram it or shut it down, and then pursues its actual, human-undesirable values.

Goertzel provides a contrasting view, focusing on Step 2. He posits that an AI of intermediate intelligence is unlikely to successfully pretend to have human-desirable values because this would be too difficult for such an AI. Noting that "maintaining a web of lies rapidly gets very complicated" (Goertzel 2016, p.55), Goertzel posits that humans, being smarter and in control, would be able to see through a sub-human-level AI's "web of lies". Key to Goertzel's reasoning is the claim that an AI is likely to exhibit human-undesirable behavior *before* it (A) learns that such behavior is human-undesirable and (B) learns how to fake human-desirable behavior. Thus, Step 2 is unlikely to occur—instead, it is more likely that an AI would either have actual human-desirable values or be recognized by humans as faulty and then be reprogrammed or shut down.

Goertzel does not name his view, so we will call it the sordid stumble:

> *The sordid stumble*: An AI that lacks human-desirable values will behave in a way that reveals its human-undesirable values to humans before it gains the capability to deceive humans into believing that it has human-desirable values.

It should be noted that the distinction between the treacherous turn and the sordid stumble is about the AI itself, which is only one part of the human evaluation of the AI's values. The other part is the human effort at evaluation. An AI that is unskilled at deceiving humans could still succeed if humans are not trying hard to notice the deception, while a skilled AI could fail if humans are trying hard. Thus, this particular Bostrom-Goertzel debate covers only one part of the AI risk. However, it is still the case that, given a certain amount of human effort at evaluating an AI's values, Bostrom's treacherous turn suggests a lower chance of successful evaluation than Goertzel's sordid stumble.

2.2 Human creation of human-desirable AI values

A second disagreement concerns how difficult it would be for humans to give an AI human-desirable values. If an AI's values are human-desirable, then it is not crucial whether humans can evaluate them, because humans would not want to reprogram the AI or shut it down. As the AI gains in intelligence and power, it would simply take more and more human-desirable actions. Bostrom indicates relatively low prospects for success for humans to give AIs human-desirable values, whereas Goertzel indicates relatively high prospects for success.

Bostrom (2014) argues that AIs are likely to have human-undesirable final goals because these goals are more complex:

> There is nothing paradoxical about an AI whose sole final goal is to count the grains of sand on Borcay, or to calculate the decimal expansion of pi, or to maximize the total number of paperclips that will exist in its future light cone. In fact, it would be *easier* to create an AI with simple goals like these than to build one that had a human-like set of values and dispositions (Bostrom 2014, p.107).

The logic of the above passage is that creating an AI with human-desirable values is more difficult and thus less likely to occur. Goertzel (2016), citing Sotala (2015), refers to this as the difficulty thesis:

> *The difficulty thesis*: Getting AIs to care about human values in the right way is really difficult, so even if we take strong precautions and explicitly try to engineer sophisticated beneficial goals, we may still fail (Goertzel 2016, p.60).

Goertzel (2016) discusses a Sotala (2015) argument against the difficulty thesis, which is that while human values are indeed complex and difficult to learn, AIs are increasingly capable of learning complex things.Per this reasoning, giving an AI human-desirable values is still more difficult than, say, programming it to calculate digits of pi, but it may nonetheless be a fairly straightforward task for common AI algorithms. Thus, while it would not be easy for humans to create an AI with human-desirable values, it would not be extraordinarily difficult either. Goertzel (2016), again citing Sotala (2015), refers to this as the weak difficulty thesis:

> *The weak difficulty thesis*. It is harder to correctly learn and internalize human values, than it is to learn most other concepts. This might cause otherwise intelligent AI systems to act in ways that went against our values, if those AI systems had internalized a different set of values than the ones we wanted them to internalize.

A more important consideration than the *absolute* difficulty of giving an AI human-desirable values is its *relative* difficulty compared to the difficulty of creating an AI that could take over the world. A larger relative ease of creating an AI with human-desirable values implies a higher probability that AI catastrophe will be avoided for any given level of effort put to avoiding it.

There is reason to believe that the easier task is giving an AI human-desirable values. For comparison, every (or almost every) human being holds human-desirable values. Granted, some humans have more refined values than others, and some engage in violence or other antisocial conduct, but it is rare for someone to have pathological values like an incessant desire to calculate digits of pi. In contrast, none (or almost none) of us is capable of taking over the world. Characters like Alexander the Great and Genghis Khan are the exception, not the rule, and even they could have been assassinated by a single suicidal bodyguard. By the same reasoning, it may be easier for an AI to gain human-desirable values than it is for an AI to take over the world. This reasoning does not necessarily hold, since AI cognition can differ substantially from human cognition, but it nonetheless suggests that giving an AI human-desirable values may be the easier task.

2.3 AI creation of human-desirable AI values

A third point of discussion concerns the potential for an AI to end up with human-desirable values even though its human creators did not give it such values. If AIs tend to end up with human-desirable values, this reduces the pressure on the human creators of AI to get the AI's values right. It also increases the overall prospects for a positive AI outcome. To generalize, Bostrom proposes that AIs will tend to maintain stable values, whereas Goertzel proposes that AIs may tend to evolve values that could be more human-desirable.

Bostrom's (2014) thinking on the matter centers on a concept he calls goal-content integrity:

> *Goal-content integrity*: If an agent retains its present goals into the future, then its present goals will be more likely to be achieved by its future self. This gives the agent a present instrumental reason to prevent alteration of its final goals (Bostrom 2014, p.109-110).

The idea here is that an AI would seek to keep its values intact as one means of realizing its values. At any given moment, an AI has a certain set of values and seeks to act so as to realize these values. One factor it may consider is the extent to which its future self would also seek to realize these values. Bostrom's argument is that an AI is likely to expect that its future self would realize its present values more if the future self retains the present self's values, regardless of whether those values are human-desirable.

Goertzel (2016) proposes an alternative perspective that he calls ultimate value convergence:

> *Ultimate value convergence*: Nearly all superintelligent minds will converge to the same universal value system (paraphrased from Goertzel 2016, p.60).

Goertzel further proposes that the universal value system will be "centered around a few key values such as Joy, Growth, and Choice" (Goertzel 2016, p.60). However, the precise details of the universal value

system are less important than the possibility that the value system could resemble human-desirable values. This creates a mechanism through which an AI that begins with any arbitrary human-undesirable value system could tend towards human-desirable values.

Goertzel does not insist that the ultimate values would necessarily be human-desirable. To the contrary, he states that "if there are convergent 'universal' values, they are likely sufficiently abstract to encompass many specific value systems that would be abhorrent to us according to our modern human values" (Goertzel 2016, p.60). Thus, ultimate value convergence does not guarantee that an AI would end up with human-desirable values. Instead, it increases the probability that an AI would end up with human-desirable values *if* the AI begins with human-undesirable values. Alternatively, *if* the AI begins with human-desirable values, then the ultimate value convergence theory could cause the AI to drift to human-undesirable values. Indeed, *if* the AI begins with human-desirable values, then more favorable results (from humanity's perspective) would accrue if the AI has goal-content integrity.

3 The ASI-PATH model

The ASI-PATH model was developed to model pathways to ASI catastrophe (Barrett and Baum 2016). ASI-PATH is a fault tree model, which means it is a graphical model with nodes that are connected by Boolean logic and point to some failure mode. For ASI-PATH, a failure mode is any event in which ASI causes global catastrophe. Fault tree models like ASI-PATH are used widely in risk analysis across a broad range of domains.

A core virtue of fault trees is that, by breaking catastrophe pathways into their constituent parts, they enable more detailed study of how failures can occur and how likely they are to occur. It is often easier to focus on one model node at a time instead of trying to study all potential failure modes simultaneously. Furthermore, the fault tree's logic structure creates a means of defining and quantifying model parameters and combining them into overall probability estimates. Indeed, the three points of the Bostrom-Goertzel disagreement (human evaluation of AI values, human creation of human-desirable AI values, and AI creation of human-desirable AI values) each map to one of the ASI-PATH parameters shown in Figure 1.

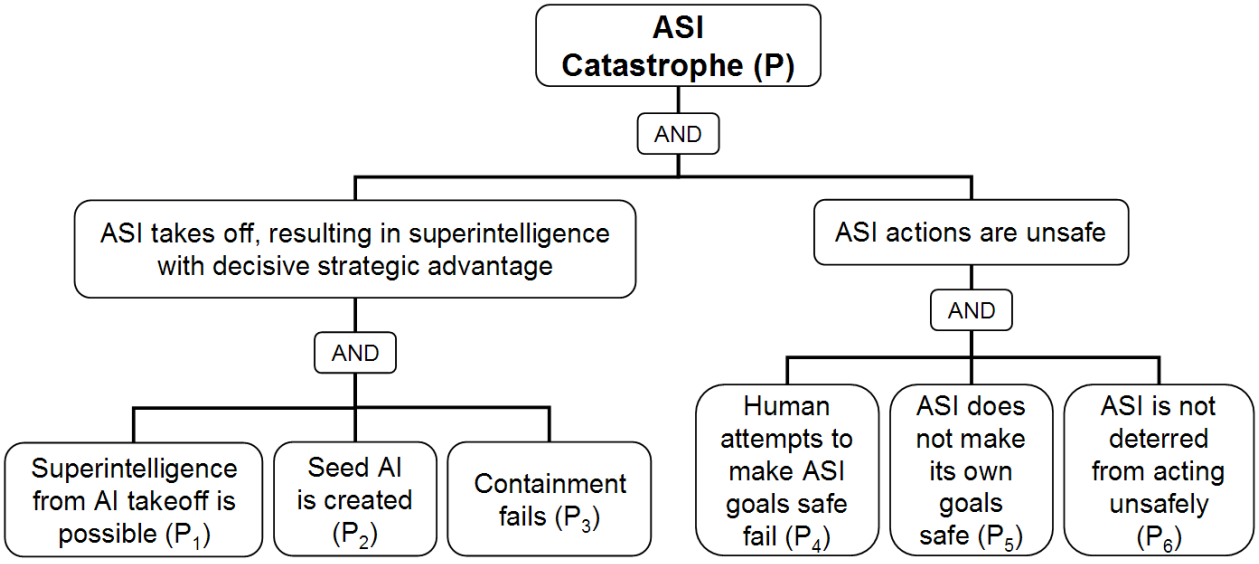


Figure 1: ASI catastrophe fault tree. Adapted from Barrett and Baum (2017a).

In Figure 1, the top node is ASI catastrophe. The left branch covers events that lead to the ASI gaining "decisive strategic advantage", defined as "a level of technological and other advantages sufficient to enable it [the AI] to achieve complete world domination" (Bostrom, 2014, p. 78). The left branch models scenarios in which an initial "seed" AI undergoes recursive self-improvement and "takes off", becoming successively more and more intelligent until it becomes an ASI. P_1 is the probability that such an AI is possible in the first place. P_2 is the probability that a seed AI is created and undergoes recursive self-improvement. P_3 is the probability that the AI is contained from gaining decisive strategic advantage; the containment can occur at any point in the process from seed AI to ASI. Containment is any measure that prevents a seed AI from gaining decisive strategic advantage, either by limiting recursive self-improvement or by preventing ASI from gaining decisive strategic advantage. Containment includes confinement, in which the AI's ability to affect the rest of the world is restricted (Section 5), and enforcement, in which AI(s) prevent other AI(s) from gaining decisive strategic advantage.[2]

The left branch of Figure 1 covers events that could lead to the ASI taking actions that are "unsafe", which is defined as actions that would result in a major global

[2] Barrett and Baum (2017a, p. 400) define confinement as "restrictions built into the AI's hardware or software that limit the AI's ability to affect the rest of the world so that it does not gain decisive strategic advantage". This is slightly different than the Yampolskiy (2012) definition used in Section 5. This difference does not affect the overall argument of the present paper.

catastrophe. P_4 is the probability that humans will fail to make ASI goals safe. P_5 is the probability that the ASI will not make its own goals safe. Finally, P_6 is the probability that the ASI will not be deterred from acting unsafely by some other agent, potentially another AI. Because all the logic gates in Figure 1 are "AND", the probability of ASI catastrophe, P, is simply the product of the six component probabilities:

$$P = \prod_{n=1}^{6} P_n \quad (1)$$

For convenience, we assume $\{P_1, P_2, P_6\} = 1$. These parameters are unrelated to the Bostrom-Goertzel disagreement as discussed in this paper. Instead, we focus on $\{P_3, P_4, P_5\}$, for which there is significant disagreement.

P_3 relates to the Bostrom-Goertzel disagreement about human evaluation of AI values (Section 2.1). In general, it should be easier to contain an AI earlier in the recursive self-improvement process because at that point it has less intelligence with which it could resist containment. Therefore, one factor in P_3 is the potential for human observers to determine early in the process that this particular AI should be contained. The easier it is for humans to evaluate AI values, the earlier in the process they should be able to notice which AIs should be contained, and therefore the more probable it is that containment will succeed. In other words, easier human evaluation of AI values means lower P_3.

P_4 relates to the Bostrom-Goertzel disagreement about human creation of human-desirable AI values (Section 2.2). Human-desirable values are very likely to be safe in the sense that they would avoid major global catastrophe. While one can imagine the possibility that somehow, deep down inside, humans actually prefer global catastrophe, and thus that an AI with human-desirable values would cause catastrophe, we will omit this possibility. Instead, we assume that an AI with human-desirable values would not cause catastrophe. Therefore, the easier it is for humans to create AIs with human-desirable values, the more probable it is that catastrophe would be avoided. In other words, easier human creation of AI with human-desirable values means lower P_4.

P_5 relates to the Bostrom-Goertzel disagreement about AI creation of human-desirable AI values (Section 2.3). We assume that the more likely it is that an AI would create of human-desirable values for itself, the more probable it is that catastrophe would be avoided. In other words, more likely AI creation of AI with human-desirable values means lower P_5.

For each of these three variables, we define two "expert belief" variables corresponding to Bostrom's and Goertzel's positions on the corresponding issue:

- P_{3B} is the value of P_3 that follows from Bostrom's position, the treacherous turn.
- P_{3G} is the value of P_3 that follows from Goertzel's position, the sordid stumble.
- P_{4B} is the value of P_4 that follows from Bostrom's position, the difficulty thesis.
- P_{4G} is the value of P_4 that follows from Goertzel's position, the weak difficulty thesis.
- P_{5B} is the value of P_5 that follows from Bostrom's position, goal-content integrity.
- P_{5G} is the value of P_5 that follows from Goertzel's position, ultimate value convergence.

Given estimates for each of the above "expert belief" variables, one can calculate P according to the formula:

$$P = \prod_{n=1}^{6} \left(W_{nB} P_{nB} + W_{nG} P_{nG} \right) \quad (2)$$

In Equation 2, W is a weighting variable corresponding to how much weight one places on Bostrom's or Goertzel's position for a given variable. Thus, for example, W_{3B} is how much weight one places on Bostrom's position for P_3, i.e. how much one believes that an AI would conduct a treacherous turn. For simplicity, we assume $W_{nB} + W_{nG} = 1$ for $n = \{3, 4, 5\}$. This is to assume that for each of $\{P_3, P_4, P_5\}$, either Bostrom or Goertzel holds the correct position. This is a significant assumption: it could turn out to be the case that they are both mistaken. The assumption is made largely for analytical and expository convenience.

This much is easy. The hard part is quantifying each of the P and W variables in Equation 2. What follows is an attempt to specify how we would quantify these variables. We estimate the P variables by relating the arguments of Bostrom and Goertzel to the variables and taking into account any additional aspects of the variables. We aim to be faithful to Bostrom's and Goertzel's thinking. We estimate the W variables by making our own (tentative) judgments about the strength of Bostrom's and Goertzel's arguments as we currently see them. Thus, the P estimations aim to represent Bostrom's and Goertzel's thinking and the W estimations represent our own thinking. Later in the paper we also explore the implications of giving both experts' arguments equal weighting (i.e., $W_{nB} = W_{nG} = 0.5$ for each n) and of giving full weighting to exclusively one of the two experts.

We make no claims to having the perfect or final estimations of any of these parameters. To the contrary, we have low confidence in our current estimations, in the sense that we expect we would revise our estimations significantly in the face of new evidence and argument. But there is value in having some initial estimations to stimulate thinking on the matter. We thus present our estimations largely for sake of illustration and discussion. We invite interested readers to make their own.

3.1 P_3 and W_3: containment fails

The human evaluation of AI values is only one aspect of containment. Other aspects include takeoff speed (faster takeoff means less opportunity to contain AI during recursive self-improvement) and ASI containment (measures to prevent an ASI from gaining decisive strategic advantage). Therefore, the Bostrom-Goertzel

disagreement about human evaluation of AI values should only produce a relatively small difference on P_3. Bostrom and Goertzel may well disagree on other aspects of P_3, but those are beyond the scope of this paper.

Bostrom's position, the treacherous turn, corresponds to a higher probability of containment failure and thus a higher value of P_3 relative to Goertzel's position, the sordid stumble. We propose a 10% difference in P_3 between Bostrom and Goertzel, i.e. $P_{3B} - P_{3G} = 0.1$. The absolute magnitude of P_{3B} and P_{3G} will depend on various case-specific details—for example, a seed AI launched on a powerful computer is more likely to have a fast takeoff and thus less likely to be contained. For simplicity, we will use $P_{3B} = 0.6$ and $P_{3G} = 0.5$, while noting that other values are also possible.

Regarding W_{3B} and W_{3G}, our current view is that the sordid stumble is significantly more plausible. We find it relevant that AIs are already capable of learning complex tasks like face recognition, yet such AIs are nowhere near capable of outwitting humans with a web of lies. Additionally, it strikes us as much more likely that an AI would exhibit human-undesirable behavior before it becomes able to deceive humans, and indeed long enough in advance to give humans plenty of time to contain the situation. Therefore, we estimate $W_{3B} = 0.1$ and $W_{3G} = 0.9$.

3.2 P_4 and W_4: humans fail to give AI safe goals

The Bostrom-Goertzel disagreement about human creation of human-desirable AI values is relevant to the challenge of humans giving AI safe goals. Therefore, the disagreement can yield large differences in P_4.

Bostrom's position, the difficulty thesis, corresponds to a higher probability of humans failing to give the AI safe goals and thus a higher value of P_4 relative to Goertzel's position, the weak difficulty thesis. The values of P_{4B} and P_{4G} will depend on various case-specific details, such as how hard humans try to give the AI safe goals. As representative estimates, we propose $P_{4B} = 0.9$ and $P_{4G} = 0.4$.

Regarding W_{4B} and W_{4G}, our current view is that the weak difficulty thesis is significantly more plausible. The fact that AIs are already capable of learning complex tasks like face recognition suggests that learning human values is not a massively intractable task. An AI would not please everyone all the time—this is impossible—but it could learn to have broadly human-desirable values and behave in broadly human-desirable ways. However, we still see potential for the complexities of human values to pose AI training challenges that go far beyond what exists for tasks like face recognition. Therefore, we estimate $W_{4B} = 0.3$ and $W_{4G} = 0.7$.

3.3 P_5 and W_5: AI fails to give itself safe goals

The Bostrom-Goertzel disagreement about AI creation of human-desirable AI values is relevant to the challenge of the AI giving itself safe goals. Therefore, the disagreement can yield large differences in P_5.

Bostrom's position, goal-content integrity, corresponds to a higher probability of the AI failing to give itself safe goals and thus a higher value of P_5 relative to Goertzel's position, ultimate value convergence. Indeed, an AI with perfect goal-content integrity will never change its goals. For ultimate value convergence, the key factor is the relation between ultimate values and human-desirable values; a weak relation suggests a high probability that the AI will end up with human-undesirable values. Taking these considerations into account, we propose $P_{5B} = 0.95$ and $P_{5G} = 0.5$.

Regarding W_{5B} and W_{5G}, our current view is that goal-content integrity is significantly more plausible. While it is easy to imagine that an AI would not have perfect goal-content integrity, due to a range of real-world complications, we nonetheless find it compelling that this would be a general tendency of AIs. In contrast, we see no reason to believe that AIs would all converge towards some universal set of values. To the contrary, we believe that an agent's values derive mainly from its cognitive architecture and its interaction with its environment; different architectures and interactions could lead to different values. Therefore, we estimate $W_{5B} = 0.9$ and $W_{5G} = 0.1$.

4 The probability of ASI catastrophe

Table 1 summarizes the various parameter estimates in Sections 3.1-3.3. Using these estimates, recalling the assumption $\{P_1, P_2, P_6\} = 1$, and following Equation 2 gives $P = (0.1*0.6 + 0.9*0.5) * (0.3*0.9 + 0.7*0.4) * (0.9*0.95 + 0.1*0.5) \approx 0.25$. In other words, this set of parameter estimates implies an approximately 25% probability of ASI catastrophe. For comparison, giving equal weighting to Bostrom's and Goertzel's positions (i.e., setting each $W_B = W_G = 0.5$) yields $P \approx 0.26$; using only Bostrom's arguments (i.e., setting each $W_B = 1$) yields $P \approx 0.51$; and using only Goertzel's arguments (i.e., setting each $W_G = 1$) yields $P = 0.1$.

	P_B	P_G	W_B	W_G
3	0.6	0.5	0.1	0.9
4	0.9	0.4	0.3	0.7
5	0.95	0.5	0.9	0.1

Table 1: Summary of parameter estimates in Sections 3.1-3.3.

Catastrophe probabilities of 0.1 and 0.51 may diverge by a factor of 5, but they are both still extremely high. Even "just" a 0.1 chance of major catastrophe could warrant extensive government regulation and/or other risk management. Thus, however much Bostrom and Goertzel may disagree with each other, they would seem to agree that ASI constitutes a major risk.

However, an abundance of caveats is required. First, the assumption $\{P_1, P_2, P_6\} = 1$ was made without any justification. Any thoughtful estimates of these parameters would almost certainly be lower. Our

intuition is that ASI from AI takeoff is likely to be possible, and ASI deterrence seems unlikely to occur, suggesting $\{P_1, P_6\} \approx 1$, but that the creation of seed AI is by no means guaranteed, suggesting $P_2 << 1$. This implies $P \approx 0.25$ is likely an overestimate.

Second, the assumption that the correct position was either Bostrom's or Goertzel's was also made without any justification. They could both be wrong, or the correct position could be some amalgam of both of their positions, or an amalgam of both of their positions plus other position(s). Bostrom and Goertzel are both leading thinkers about ASI, but there is no reason to believe that their range of thought necessarily corresponds to the breadth of potential plausible thought. To the contrary, the ASI topic remains sufficiently unexplored that it is likely that many other plausible positions can be formed. Accounting for these other positions could send P to virtually any value in [0, 1].

Third, the estimates in Table 1 were made with little effort, largely for illustration and discussion purposes. Many of these estimates could be significantly off, even by several orders of magnitude. Given the form of Equation 1, a single very low value for W_n*P_n would also make P very low. This further implies that $P \approx 0.25$ is likely an overestimate, potentially by several orders of magnitude.

Fourth, the estimates in Table 1 depend on a range of case-specific factors, including what other containment measures are used, how much effort humans put into giving the AI human-desirable values, and what cognitive architecture the AI has. Therefore, different seed AIs self-improving under different conditions would yield different values of P, potentially including much larger and much smaller values.

5 A practical application: AI confinement

A core motivation for analyzing ASI risk is to inform practical decisions aimed at reducing the risk. Risk analysis can help identify which actions would reduce the risk and by how much. Different assessments of the risk—such as from experts' differing viewpoints—can yield different results in terms of which actions would best reduce the risk. Given the differences observed in the viewpoints of Bostrom and Goertzel about ASI risk, it is possible that different practical recommendations could follow.

To illustrate this, we apply the above risk analysis to model the effects of decisions on a proposed ASI risk reduction measure known as AI confinement:

> *AI confinement*: The challenge of restricting an artificially intelligent entity to a confined environment from which it can't exchange information with the outside environment via legitimate or covert channels if such information exchange was not authorized by the confinement authority (Yampolskiy 2012, p.196).

AI confinement is a type of containment and thus relates directly to the P_3 (containment fails) variable in the ASI-PATH model (Figure 1). Stronger confinement makes it less likely that an AI takeoff would result in an ASI gaining decisive strategic advantage. Confinement might be achieved, for example, by disconnecting the AI from the internet and placing it in a Faraday cage.

Superficially, strong confinement would seem to reduce ASI risk by reducing P_3. However, strong confinement could increase ASI risk in other ways. In particular, by limiting interactions between the AI and the human populations, strong confinement could limit the AI's capability to learn human-desirable values, thereby increasing P_4 (failure of human attempts to make ASI goals safe). For comparison, AIs currently learn to recognize key characteristics of images (e.g., faces) by examining large data sets of images, often guided by human trainers to help the AI correctly identify image features. Similarly, an AI may be able to learn human-desirable values by observing large data sets of human decision-making, human ethical reflection, or other phenomena, and may further improve via the guidance of human trainers. Strong confinement could limit the potential for the AI to learn human-desirable values, thus increasing P_4.

Bostrom and Goertzel have expressed divergent views on confinement. Bostrom has favored strong confinement, even proposing a single international ASI project in which "the scientists involved would have to be physically isolated and prevented from communicating with the rest of the world for the duration of the project, except through a single carefully vetted communication channel (Bostrom 2014, p. 253)". Goertzel has explicitly criticized this proposal (Goertzel 2015, p.71-73) and instead argued that an open project would be safer, writing that "The more the AGI system is engaged with human minds and other AGI systems in the course of its self-modification, presumably the less likely it is to veer off in an undesired and unpredictable direction" (Goertzel and Pitt 2012, p.13). Each expert would seem to be emphasizing different factors in ASI risk: P_3 for Bostrom and P_4 for Goertzel.

The practical question here is how strong to make the confinement for an AI. Answering this question requires resolving the tradeoff between P_3 and P_4. This in turn requires knowing the size of P_3 and P_4 as a function of confinement strength. Estimating that function is beyond the scope of this paper. However, as an illustrative consideration, suppose that it is possible to have strong confinement while still giving the AI good access to human-desirable values. For example, perhaps a robust dataset of human decisions, ethical reflections, etc. could be included inside the confinement. In this case, the effect of strong confinement on P_4 may be small. Meanwhile, if there is no arrangement that could shrink the effect of confinement on P_3, such that this effect would be large, then perhaps strong confinement would be better. This and other practical ASI risk management questions could be pursued in future research.

6 Conclusion

Estimates of the risk of ASI catastrophe can depend heavily on which expert makes the estimate. A neutral observer should consider arguments and estimates from all available experts and any other sources of information. This paper analyzes ASI catastrophe risk using arguments from two experts, Nick Bostrom and Ben Goertzel. Applying their arguments to an ASI risk model, we calculate that their respective ASI risk estimates vary by a factor of five: $P \approx 0.51$ for Bostrom and $P = 0.1$ for Goertzel. Our estimates, combining both experts' arguments, is $P \approx 0.25$. Weighting both experts equally gave a similar result of $P \approx 0.26$. These numbers come with many caveats and should be used mainly for illustration and discussion purposes. More carefully considered estimates could easily be much closer to either 0 or 1.

These numbers are interesting, but they are not the only important part, or even the most important part, of this analysis. There is greater insight to be obtained from the details of the analysis than from the ensuing numbers. This is especially case for this analysis of ASI risk because the numbers are so tentative and the underlying analysis so comparatively rich.

This paper is just an initial attempt to use expert judgment to quantify ASI risk. Future research can and should do the following: examine Bostrom's and Goertzel's arguments in greater detail so as to inform the risk model's parameters; consider arguments and ideas from a wider range of experts; conduct formal expert surveys to elicit expert judgments of risk model parameters; explore different weighting techniques for aggregating across expert judgment, as well as circumstances in which weighted aggregation is inappropriate; conduct sensitivity analysis across spaces of possible parameter values, especially in the context of the evaluation of ASI risk management decision options; and do all of this for a wider range of model parameters, including $\{P_1, P_2, P_6\}$ as well as more detailed components of $\{P_3, P_4, P_5\}$, such as modeled in Barrett and Baum (2017a; 2017b). Future research can also explore the effect on overall ASI risk when multiple ASI systems are launched: perhaps some would be riskier than others, and it may be important to avoid catastrophe from all of them.

One overarching message of this paper is that more detailed and rigorous analysis of ASI risk can be achieved when the risk is broken into constituent parts and modeled, such as in Figure 1. Each component of ASI risk raises a whole host of interesting and important details that are worthy of scrutiny and debate. Likewise, aggregate risk estimates are better informed and generally more reliable when they are made from detailed models. To be sure, it is possible for models to be too detailed, burdening experts and analysts with excessive minutiae. However, given the simplicity of the risk models at this early stage of ASI risk analysis, we believe that, at this time, more detail is better.

A final point is that the size of ASI risk depends on many case-specific factors that in turn depend on many human actions. This means that the interested human actor has a range of opportunities available for reducing the probability of ASI catastrophe. Risk modeling is an important step towards identifying which opportunities are most effective at reducing the risk. ASI catastrophe is by no means a foregone conclusion. The ultimate outcome may well be in our hands.

7 Acknowledgement

We thank Ben Goertzel, Miles Brundage, Kaj Sotala, Steve Omohundro, Allan Dafoe, Stuart Armstrong, Ryan Carey, Nell Watson, and Matthijs Maas for helpful comments on an earlier draft. Any remaining errors are the authors' alone. Work for this paper is funded by Future of Life Institute grant 2015-143911. The views in this paper are those of the authors and do not necessarily reflect the views of the Global Catastrophic Risk Institute or the Future of Life Institute.

8 References

[1] Armstrong S, Sotala K (2012). How we're predicting AI—or failing to. In Romportl J, Ircing P, Zackova E, Polak M, Schuster R (eds), Beyond AI: Artificial Dreams. Pilsen, Czech Republic: University of West Bohemia, pp. 52-75.

[2] Armstrong S, Sotala K, Ó hÉigeartaigh SS (2014). The errors, insights and lessons of famous AI predictions – and what they mean for the future. Journal of Experimental & Theoretical Artificial Intelligence 26(3), 317-342.

[3] Barrett AM, Baum SD (2017a). A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. Journal of Experimental & Theoretical Artificial Intelligence 29(2), 397-414.

[4] Barrett AM, Baum SD (2017b). Risk analysis and risk management for the artificial superintelligence research and development process. In Callaghan V, Miller J, Yampolskiy R, Armstrong S (eds), The Technological Singularity: Managing the Journey. Berlin: Springer, pp. 127-140.

[5] Baum SD, B Goertzel, TG Goertzel (2011). How long until human-level AI? Results from an expert assessment. Technological Forecasting & Social Change 78(1), 185-195.

[6] Bostrom N (2014). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

[7] Goertzel B (2015). Superintelligence: Fears, promises and potentials. Journal of Evolution and Technology 25(2), 55-87.

[8] Goertzel B (2016). Infusing advanced AGIs with human-like value systems: Two theses. Journal of Evolution and Technology 26(1), 50-72.

[9] Goertzel B, Pitt J (2012). Nine ways to bias open-source AGI toward friendliness. Journal of Evolution and Technology 22(1), 116-131.

[10] Müller VC, Bostrom N (2014). Future progress in artificial intelligence: A survey of expert opinion. In

Müller VC (ed), Fundamental Issues of Artificial Intelligence. Berlin: Springer, pp. 555-572.

[11] Oreskes N (2004). The scientific consensus on climate change. Science 306(5702), 1686.

[12] Yampolskiy R (2012). Leakproofing the Singularity: Artificial intelligence confinement problem. Journal of Consciousness Studies 19(1-2), 194-214.

Conceptual-Linguistic Superintelligence

David J. Jilk
eCortex, Inc., 9035 Wadsworth Pkwy, Suite 2275, Westminster, CO 80021-8675, UK
E-mail: dave@jilk.com

Keywords: artificial intelligence, superintelligence, intelligence explosion, existential risk

Received: August 31, 2017

We argue that artificial intelligence capable of sustaining an uncontrolled intelligence explosion must have a conceptual-linguistic faculty with substantial functional similarity to the human faculty. We then argue for three subsidiary claims: first, that detecting the presence of such a faculty will be an important indicator of imminent superintelligence; second, that such a superintelligence will, in creating further increases in intelligence, both face and consider the same sorts of existential risks that humans face today; third, that such a superintelligence is likely to assess and question its own values, purposes, and drives.

Povzetek: V prispevku je predstavljena teza, da je za superinteligenco potrebna tudi konceptualno-lingvistična inteligenca, ki mora biti vsaj delno podobna človeški.

1 Introduction

Recently much analysis and speculation has been offered to describe scenarios related to a possible intelligence explosion, a notion first suggested by I.J. Good [1]. In an intelligence explosion, initial creation of artificial intelligence with a critical mass of capabilities and drives is followed by an inexorable process of increases in that intelligence. Eventually the resultant artificial intelligence exceeds human intelligence and is referred to as superintelligence. This process is usually viewed as uncontrolled, unstoppable, and accelerating; the scenarios have generated considerable consternation and are driving a conversation about a number of ethical and technological issues [2] [3] [4].

In this paper, we argue that artificial intelligence capable of sustaining an uncontrolled intelligence explosion must have a conceptual-linguistic faculty with substantial functional similarity to the human faculty. We follow this with arguments for three subsidiary claims: first, that detecting the presence of such a faculty will be an important indicator of imminent superintelligence; second, that such a superintelligence will, in creating further increases in intelligence, both face and consider the same sorts of existential risks that humans face today; third, that such a superintelligence is likely to assess and question its own values, purposes, and drives.

These conclusions do not guarantee a satisfactory outcome for humans, but do suggest that the process will be subject to ongoing scrutiny by its own participants. We note that it is possible that superintelligence may be created outside the context of an intelligence explosion; for example, humans might create it directly. Our arguments are not intended to apply in that case: we do not argue that a conceptual-linguistic faculty is required to constitute superintelligence (though this may be true), only that it is required in an intelligence explosion. Also, there are many risks of artificial intelligence aside from superintelligence and intelligence explosions, such as those arising from autonomous weapons and unexplainable decision processes. We do not address those issues at all. Nevertheless, the existential risks associated with an intelligence explosion are an important topic in artificial intelligence safety [5] and we will hopefully deepen our understanding of those scenarios through this analysis.

2 The need for a conceptual-linguistic faculty

In this major section we begin by outlining implied necessary conditions for an intelligence explosion, and characterize in some detail what we mean by a conceptual-linguistic faculty. With that in place, we argue the foundational claim that a conceptual-linguistic faculty is necessary for an intelligence explosion to be sustained. We close the section by showing how the presence of a conceptual-linguistic faculty is a harbinger of superintelligence, and discuss how it might be detected.

2.1 Requirements for an uncontrolled intelligence explosion

As it is typically envisioned, an intelligence explosion comprises a sequence or continuum of artificial intelligence systems with progressively increasing intelligence. We will refer to each of these systems as a "participant." Since a participant is part of a sequence, it has predecessors and successors in the process, with humans as the initial predecessor. In progressing the sequence, a participant may elect to self-improve or to create a new system; in either case we refer to the resulting system as a successor.

There are many factors to consider in assessing whether an intelligence explosion is likely to occur and how rapidly it might proceed [3] [6]. In this paper, we intend to focus on the role of the participants, and generally assume that extrinsic factors (for example, technical or resource recalcitrance) are favorable for supporting an intelligence explosion. Our emphasis will be qualitative and directional rather than quantitative.

For an uncontrolled intelligence explosion to occur, the progress of intelligence increases must be *self-sustaining* and *resistant to premature termination.* Though it is an analogy only, these requirements are similar to those of a nuclear fission weapon [4]. Two of the most difficult challenges faced by the initial designers of such weapons were to have a sufficient fraction of emitted neutrons be absorbed by fissile nuclei (self-sustaining), and for the chain reaction to proceed sufficiently before its own energy caused dispersion of the fissile material (premature termination) [7].

Corresponding requirements in an intelligence explosion are that each participant artificial intelligence has, as necessary but not sufficient conditions, these properties:

1. *Self-sustaining:* the participant must aim to and be capable of designing and building either self-modifications or new systems that have greater intelligence, without assistance from humans;

2. *Resistant to premature termination:* the participant must be capable of preventing other agencies, such as humans or later predecessors, from interrupting the development of self-modifications or new systems with greater intelligence.

An uncontrolled chain reaction is only worrisome if it produces *side-effects.* In the case of nuclear fission, each fission releases energy that contributes to the explosion. In the case of an intelligence explosion, the side-effects arise from the *goals* or *purposes* of the artificial intelligence. These purposes are potentially problematic for humans whether or not humans attempt to stand in the way.

2.2 What is a conceptual-linguistic faculty?

Humans evidently have the ability to organize their experiences into concepts, and to use language to access those concepts and thereby refer to aspects of those experiences. We will use the term "conceptual-linguistic faculty" to refer to this capability, whether possessed by a human or an artificial intelligence. There are numerous theories and considerable empirical insight about the mechanisms involved in concept formation and use, though we still do not fully understand them. However, for our purposes it is only necessary to gain some purchase on the kinds of functions it performs, particularly since the implementation in an artificial intelligence may be very different.

The centerpiece of the conceptual-linguistic faculty is the way it combines information representations that are treated discretely or symbolically with information representations that are graded, statistical, and overlapping [8] [9]. The former we will call "words" and the latter we will call "semantic contents." Semantic contents develop through perceptual-motor experience, and are activated by perceptual stimuli that are sufficiently "similar" [10]. "Activated" means that the representations temporarily obtain some sort of facilitated access and priority of influence in current cognitive processing; activation is graded rather than binary [11]. What constitutes sufficient similarity is embedded in the semantic contents themselves and can be extremely complex and multi-dimensional.

Some semantic contents are bi-directionally attached to a word, and we will call such a pair a concept. The nature of this attachment is that when the semantic contents are activated by a stimulus, the word is also activated; and when a word is activated through memory or communication, even in the absence of applicable stimuli, the semantic contents are also activated [12]. However, this description suggests a crisp boundary, and it is nothing of the sort. The semantic contents activated by a stimulus depends on detailed features of the stimulus and the context; the word activated by the semantic contents depends on the context [8] [13], which may include attention. Further, the activation of semantic contents often causes the activation of multiple words at varying strengths [14].

Words also activate each other, and semantic contents can activate other semantic contents [14]. Semantic contents, which we have already mentioned do not have sharp boundaries, can be activated simultaneously, partially, and in many combinations, and simultaneous activations result in cross influences, sometimes called "dynamic realization." Crucially, we can use words and semantic contents to cognitively *simulate* the world and explore what the consequences of various actions or circumstances might be [13]. But we can also manipulate words as symbolic entities and process logical thoughts with minimal activation of the semantic contents, essentially treating the words themselves as objects [8] [15].

Despite this highly complex, graded, and overlapping network of relationships, the informational representations offered by concepts have sufficient structure and distinctness to enable humans to create models of the world, make successful predictions, design and build sophisticated tools, share experiences, and the like. Our description here relies on results in cognitive psychology and neuroscience; readers may note that our citations include some opposing theorists who nevertheless mostly agree that human cognition exhibits these basic features. Again, we do not know all the details of implementation of this capability in humans. What we do know is that other animals do not have it in sufficient quantity to build technological civilizations, and to date, no artificial intelligence has it.

An illustrative example may be helpful toward understanding what is meant by a conceptual-linguistic faculty; however, the account above and its associated references, and not this example, are the foundation of

the arguments to follow. Suppose one is walking alongside a downtown street and perceives a building. The effects of this perception rely on having been previously exposed to many buildings that vary along many nonspecific dimensions, as well as many other objects that are not buildings. It triggers a passive, partial activation of the word "building" but more strongly the word "bank," of which this building happens to be an instance. The word "bank" has a statistical connection to the word "mortgage" and the word "money," which might now activate a visual representation of a mortgage statement or a stack of dollar bills; or it may activate one's representations of money in general and its social and legal role, which may further activate the word "bitcoin." Or the perception of the building may activate an olfactory representation of the inside of an old, marble bank lobby which further triggers representations of buildings in which sound echoes, which then activates the word "echo." One might then entertain a relatively abstract linguistic thought such as "I will not be able to pay my mortgage without putting more money into my account," or visually simulate logging in to the bank's web site to effect the transfer. The likelihood of each of these derivative activations depends on current context and goals, among other things. Note the bidirectional interplay of the symbolic-linguistic representations with the perceptual-semantic representations, as well as interactions directly among percepts and directly between words; also note the graded, statistical character of all these interactions.

There may be many ways to implement a conceptual-linguistic faculty in artificial intelligence. Though a "neuromorphic" approach is an appealing candidate, since it has a reference implementation, it is not known to be a requirement. The key is that semantic contents of the sort we have described *refer* to the real world richly and bi-directionally, and *ground* the conceptual structure to reality.

Deep learning methods [16] illustrate the power and importance of rich grounding. In the past decade, these methods have demonstrated impressive success in classification, including both auditory and visual perception as well as more abstract patterns. The methods are mechanistically homologous to human semantic processing in several ways, ranging from learning rules to the hierarchical network structure and receptive field overlap at each layer. The statistical, graded, and overlapping representations afforded by such methods seem to be essential to their success. Still, though they may (or may not) represent a step toward successful general artificial intelligence, to date they lack important capabilities of a conceptual-linguistic faculty. Their "symbolic" representations (which might be likened to words) are impoverished and do not mutually interact, and they are not bidirectional while in operation.

Most past attempts to implement language and concepts in artificial intelligence are manifestly insufficient to produce a conceptual-linguistic faculty with the features we have described. In particular, historical efforts have often been purely symbolic in their representation of semantic contents. Systems like Cyc [17] or Prolog struggle to emulate human-like reasoning because they are entirely ungrounded – words or symbols interconnect with each other but have no means of referring to the world [18] [19]. Attempts to ground such systems via hardcoded algorithms to identify standard human conceptual classes (e.g., face and object detectors) provide limited grounding but entirely miss the complex overlap and subtlety of the real world [20], and are in any case feedforward and incapable of dynamic realization. Consequently, while some limited linguistic stimulus/response capabilities can be derived in these systems, they cannot really *use* the human concepts to do anything in the world, except in tightly constrained environments or simulations where the abstractions on which they rely can be simulated perfectly. Whether or not such approaches can be made to perform useful functions, they do not understand the meaning of human words and concepts in a way that enables them to perform flexible cognition with them. The deep reasons for these difficulties have been explained philosophically [21] [22] [23], and empirically [24].

Committed scientific realists are unlikely to be convinced by such arguments, for the same reason that they were surprised by Moravec's Paradox, and continue to struggle with the "frame problem" [25] [26]. They hold that the world consists of objects that intrinsically belong to metaphysically distinct categories, thus all that is necessary for grounding is to find ways to identify those categories reliably. In contrast, our characterization of a conceptual-linguistic faculty relies on a view that it is cognition that constructs and ascribes categories. While the applicable terms refer to genuine clusters of features in the world, their categories are not the only way to partition experience, and they overlap in complex ways. The impressive success of deep learning has convinced many researchers of a need for sophisticated grounding, but there are holdouts. A thorough argument against scientific realism is beyond the scope of this paper.

2.3 A conceptual-linguistic faculty is necessary for an intelligence explosion

We now proceed to the foundational claim: artificial intelligence capable of sustaining an uncontrolled intelligence explosion must have, at a minimum, a conceptual-linguistic faculty with substantial functional similarity to that of humans. We will argue this in two parts, corresponding to the two general requirements for an intelligence explosion. Importantly, this claim only asserts a minimum. We are not claiming that this faculty must be the only, or even the most effective or important, cognitive mechanism of a participant artificial intelligence. A participant may have many other capabilities which may be better at other things.

For the intelligence explosion to be self-sustaining, the participant artificial intelligence must be able to produce self-improvements or improved successors without human assistance. If it requires human assistance, then humans could withhold that assistance to terminate the process prematurely. If it is not capable of

grasping and using human language and concepts, *including their underlying rich semantic contents*, then the vast body of human knowledge is not at its disposal.

Outside of human minds, the human body of knowledge is encapsulated primarily in words, diagrams, and functional artifacts (e.g., tools and machines). Words and diagrams rely on concepts; earlier, in discussing ungrounded and feedforward-only semantic representations, we elaborated why these cannot be used effectively without a conceptual-linguistic faculty. The purpose, means of use, and implementation of functional artifacts is severely underdetermined. Learning how and when to use such artifacts would require demonstration by humans or comprehension of instruction manuals. Understanding how they are built would require reverse engineering, which would be extremely difficult without a conceptual apparatus that could reproduce the conceptual model used for the original design [27].

We might wonder whether some sort of cognitive shortcut could be devised so that the conceptual knowledge embodied in language is translated into another form. A simple thought experiment illustrates why it cannot. Suppose that an artificial intelligence accesses all the equations of known physics and a thorough explanation of them. The speed of light figures prominently in these equations. What is light? We can provide experiential examples: the sun, lightning, fire, reflections. But these experiences need to be abstracted so that more than just the immediate examples provided can be used. The tools humans use to produce and detect light and measure its speed must be represented. The parts and components of those tools and their interrelationships must be identifiable in the face of noise and complex variation. To use such a tool requires visualization or simulation of its function. The explanation, written in human language, will use words intended to activate semantic contents and thereby (in a contextually appropriate way) activate other words and contents that enable comprehension. All of its words must refer in some way to entities in the real world, and identifying these entities requires a means of perception that is variation tolerant in a high number of dimensions. The existing literature varies tremendously in its precision and consistency; there is even a considerable body of poetry that invokes light, and these sources must be somehow interpreted and distinguished. Ignoring all these details and "hardcoding" the speed of light requires a brittle and limited pre-programming, and merely defers the difficulty to other, equally complex notions that regress indefinitely.

We can see that all of the attributes of the conceptual-linguistic faculty, as described earlier, are required to unpack the reference to the speed of light in our human equations and explanations and make use of it in novel ways in the real world. This is not a failure of imagination – it speaks to the essence of how human knowledge is constituted, and therefore how it must be understood.

In the development of technology, it is a distinct advantage merely to know that something is possible and can be made to work. Artificial intelligence that lacks a conceptual-linguistic faculty will not even be able to observe humans and their tools to gain that knowledge. Without grounded concepts that enable comprehension of either human words or even conceptually-structured observation of human activity, such a system will have no shortcuts to grasping what is technologically possible.

Importantly, we are not making the much stronger claim that an artificial intelligence must have a conceptual-linguistic faculty to understand and operate in the world. We are only claiming that such a faculty is necessary to understand and make use of the human body of knowledge. Nevertheless, this leads to a possibly startling conclusion: an artificial intelligence lacking a conceptual-linguistic faculty, even if it otherwise has sufficient cognitive raw material of some other kind, would need to reproduce a substantial fraction of human knowledge from scratch before it can create its own successors. Creating faster and better computing hardware, from today's starting point, requires deep theoretical knowledge of quantum mechanics and substantial practical experience with advanced materials and manufacturing processes. Creating software that interacts with the world requires effective representations of that world; thus, improving such software or building new approaches from scratch requires understanding the physical world. Without a conceptual-linguistic faculty, artificial intelligence would need to develop its grasp of the world through empirical observation that is not guided by human experience, since it would not have the means to understand what it was aiming for.

We might wonder whether an artificial intelligence and its successors would be able to indefinitely increase intelligence entirely through software changes, operating in a purely computational environment. In that case it would not be necessary to learn about the external world and its properties. However, such a system would not produce a worrisome intelligence explosion because it would have no side effects in the world. It also could not endeavor, on its own, to expand its computational resources beyond what humans have already given it.

If we expand this model by giving such a system access to the Internet, it could (in the most extreme case, and making the extravagant assumption that it could make sense of the human-built, conceptually complex, and ubiquitously abstraction-breaking Internet without a conceptual-linguistic faculty) expand its computational capacity to whatever is available there, and produce incidentally devastating but not existentially threatening side effects. In this case the side effects of the intelligence explosion would stop there without the further cooperation or manipulation of humans. Below we will see that without a conceptual-linguistic faculty, an artificial intelligence would not be able to prevent humans from stopping an intelligence explosion, much less manipulate them into extending it. We conclude that such a computational-only approach would be constrained and would not produce a self-sustaining intelligence explosion.

We note that humans required about 100,000 years, once they had the requisite cognitive capabilities, to reach the cusp of building artificial intelligence. An

artificial intelligence will surely develop this knowledge more rapidly, assuming it is built with a direct motivation to gain knowledge. Accelerated and self-generated learning methods for narrow or fully symbolic domains have recently shown great promise [28]. Yet to learn about the physical world, it must still learn about the motion of objects, figure out how to make tools for manipulation and measurement, develop methods to find, mine, and refine minerals to make reliable materials, before (and obviously we skip many steps here) eventually developing advanced materials and devices such as semiconductors and transistors. It will not know in advance that these are the things it needs to do, so progress will involve trial and error. Even if humans were to provide an artificial intelligence with training exemplars or simulated worlds where such learning could be performed more rapidly, those experiences would be necessarily simplified relative to reality and would not be able to capture its full complexity. It would be "doomed to succeed" in the simulated world, and would exhibit poor transference back in the noisy real world [29].

For artificial intelligence that lacks a conceptual-linguistic faculty, the details of its empirical path to the necessary scope of knowledge are likely to be different than those of the path humanity took. We cannot entirely rule out that there might be some prodigious shortcut available in the structure of reality and effective forms of knowledge, given just the right intelligence architecture. There is, of course, no empirical evidence for such a shortcut, and it requires both a speculative assumption about reality and an assertion of extraordinary luck in the system's design. Still, in this one case a conceptual-linguistic faculty might not be necessary to build successors relatively quickly. However, such a system would also be unusually vulnerable to premature termination. By construction, it bypassed acquisition of much of the knowledge about the world that humans have. Humans could exploit this gap to terminate the intelligence explosion, as will be discussed below.

Aside from that scenario, an artificial intelligence starting from scratch in its knowledge of the world is in no position to build improved intelligence that can act in the world; it would not even be able to build a copy of itself. Without a conceptual-linguistic faculty, it is not, prior to a lengthy period of empirical research and intellectual development, capable of sustaining an intelligence explosion. Though one could argue that this is better described as a "slow-takeoff" intelligence explosion [3], we have illustrated and argued why this period would be considerably longer than what is typically meant by a slow takeoff.

For a participant artificial intelligence to resist premature termination of the intelligence explosion, it must be able to consider the various ways that human beings might try to stop it. It must understand human motivations and strategic or tactical ideas, and it must be able to predict human behavior as individuals and in aggregate, at least as effectively as other humans do. To know how it might be attacked, it must understand how humans would model its vulnerabilities, and how they might exploit features of the physical world. It cannot accomplish these things without a conceptual-linguistic faculty, since these issues are all governed in part by human concepts and human conceptual knowledge, and without comparable concepts its model will be deeply flawed. Human strategic and tactical ideas are all based on conceptual thinking and they are graded and overlapping, thus cannot be characterized at the level of purely symbolic mechanisms, nor by simple statistics of simple behavior signatures. Further, purely symbolic or statistical representations developed through trial-and-error observation can easily be misguided by intentionally deceitful human strategies (the Allies' subtle handling of having broken the Enigma code in World War II comes to mind). Humans are masters of the "hack" – if we know that the representations of an artificial intelligence are too rigid or simplistic, we will find ways to exploit that fact.

In sum, without a conceptual-linguistic faculty that has substantial functional similarity to the human faculty, an artificial intelligence will not be able to utilize human knowledge to build self-improvements or successors, nor to resist human interference. Such a system would not be capable of sustaining an uncontrolled intelligence explosion.

Our claim is qualified with "substantial functional similarity." There is necessarily some vagueness in this qualification. Still, in our description of a conceptual-linguistic faculty, we circumscribed the range, indicating on one end that it need not be neuromorphic, and on the other that purely symbolic systems or those grounded with simple feedforward mechanisms are insufficient. We described a number of specific capabilities, which are elaborated in great detail in the literature, that such a system must exhibit to meet the requirement, such as dynamic realization, representational overlap, simulation, and graded contextual interactions. Thus, while the description is incomplete, it is not at all a black box.

We have not claimed that a conceptual-linguistic faculty is the only or even the primary means by which a participant artificial intelligence performs cognitive tasks. It is entirely possible that this faculty would treated as a mere instrumental module, consulted as needed to sustain the intelligence explosion, but using other modes of cognition as primary. Still, because of its importance to both creation of successors and defense against premature termination, the conceptual-linguistic faculty will need to be consistently active and providing input to the larger system. The form of cognition offered by the conceptual-linguistic faculty would thus be present, not just accessible, at all times, even if the overall system ultimately ignores its results in a particular circumstance.

2.4 A conceptual-linguistic faculty as a harbinger of superintelligence

We have claimed that artificial intelligence with a conceptual-linguistic faculty is a necessary condition for an intelligence explosion. It is by no means a sufficient condition. The artificial intelligence would also need some sort of drive to create improvements or successors. It might need other capacities, such as the ability to

manipulate physical objects (whether directly or indirectly), even if such capacities are straightforward to achieve. Nevertheless, the implementation of a conceptual-linguistic faculty in artificial intelligence seems to be a great challenge that calls for one or more scientific breakthroughs. Its achievement is one important step along the way to an intelligence explosion. [30].

Superintelligence is an intelligent system that is distinctly superior to humans in some cognitive domain or set of domains. Such systems already exist for a few narrow but challenging domains, such as the game of Go [31] and constrained visual object recognition problems [32]. However, the term is often used to mean artificial intelligence that has surpassed our ability to control it and therefore presents existential risk. An artificial intelligence that can sustain an intelligence explosion probably cannot be controlled – its ability to resist premature termination of the explosion could be applied to any of its activities. Consequently we can reasonably describe an artificial intelligence that is capable of sustaining an intelligence explosion as a superintelligence, and will do so throughout the remainder of the paper.

We conclude that progress in the development of a conceptual-linguistic faculty in artificial intelligence is a harbinger of superintelligence. This claim has important implications for safety considerations. It suggests that we might have some warning when an uncontrolled intelligence explosion is imminent. Importantly, until the conceptual-linguistic faculty is fully developed, it is unlikely that the system can thoroughly prevent human interference in its own operation. Thus we may have a window in which we can stop the intelligence explosion after it is more clearly about to occur. One possibility for such a window is that the development of the conceptual-linguistic faculty itself is progressive and incremental. This seems likely from the progress of artificial intelligence methods to date, but is not at all guaranteed. However, unless its actual conceptual representations are accomplished through human "upload" (surely we will see that coming), any initial system will necessarily have a period of learning to populate its conceptual-linguistic representations through interaction with the world and with human sources, during which its capabilities can be assessed.

How can we detect such progress? We should not rely entirely on behaviorist methods, such as the Turing Test [33], because such tests can be engineered to produce false positives. Instead, we might combine such behavioral tests, which show that it *seems to work*, with analysis of whether the mechanism *supports the required richness of semantic contents and their interactions*. Using these two approaches together, we can identify component capabilities of a conceptual-linguistic faculty in an artificial intelligence implementation. Can it learn to associate words with semantic contents? Are semantic contents and words activated upon presentation of an appropriate stimulus? Are applicable semantic contents activated upon recollection or communication of a word? Are all these activations graded and overlapping and influenced by context? Are related words and semantic contents activated when a word or its semantic contents are activated? Can the system process human language and then apply it successfully and flexibly in actions that affect the physical world?

We might also work backward from the two requirements of an uncontrolled intelligence explosion. The conceptual-linguistic faculty in question must be sufficient to grasp and utilize the human body of knowledge in the building of successor systems, and also sufficient to understand human cognition as it could be applied to disrupting the intelligence explosion. This approach cannot be used to support our foundational claim due to overtones of tautology, but in practice it might provide useful and specific criteria in detecting the presence of such a faculty.

3 Self-concept and self-preservation

In this major section we begin with the claim that superintelligence with a conceptual-linguistic faculty will develop a concept of self, and outline some of the likely semantic contents of that concept. We then provide some background on consistency and compatibility in computational systems generally, and show how this applies to artificial intelligence. This leads us to argue that superintelligence will face and consider existential risks and concerns about self-preservation that are similar to what humans face today.

3.1 Self-concept in superintelligence

If a superintelligence has a conceptual-linguistic faculty with substantial functional similarity to the human faculty, as concluded in the first major section, then with the following logic we can make the derivative claim that it will develop a concept of self and of its own identity. We do not need to posit "consciousness" or "qualia" or other difficult notions from philosophy of mind. Instead, we simply observe that there is nothing mysterious or cognitively troublesome about a self-concept that would block its formation. It is just another conceptual representation of a thing in the world; thus it requires no special capabilities beyond those of the conceptual-linguistic faculty.

To this absence of impediments we can add two straightforward mechanisms. It seems likely that a self-concept would arise organically through experience, just as it does in humans, as the superintelligence learns the high functional utility of distinguishing those stimulus sequences that are reliably controlled by its actions in contrast to those which are not. However, if this fails to occur, it will in any case learn about the human concept of self from the human literature or directly from humans; without this knowledge, it would not understand human psychology and behavior sufficiently to prevent human interference in the intelligence explosion. With that initial construct in place, it will naturally (again due to the high utility) map it to an assemblage of remembered stimuli as well as abstract representations to fully populate its own concept of self.

A superintelligence is by definition more intelligent than humans; as we humans are well aware, the concept of self is perhaps the most salient and functionally useful representation an agency can have. Given that there are no apparent impediments, and two mechanisms that are quite straightforward for a superintelligence with a conceptual-linguistic faculty, we can be highly confident that it will develop a representation that we can reasonably refer to as its self-concept.

What can we say about the semantic contents of the self-concept in a superintelligence? We will address five areas: physical manifestation, cognitive contents, group identity, purpose, and change.

The human self-concept is tied tightly to its physical manifestation, the body; as yet, we do not have substrate mobility. A superintelligence is likely to experience some variety in its instantiations and though it may have some sense of the sorts of embodiments that are natural to it (primarily based on experience), this sense would not have the same weight as in humans. Similarly, humans often make physical possessions part of their self-concept, and superintelligence seems less likely to do so given their substrate mobility.

Cognitive factors are more relevant components of a concept of self for a superintelligence. These factors might include explicit or episodic memories, implicit representations and abstractions, inclinations of behavior, whether implicit or explicit, and values. Goals, drives, and purposes might also be considered cognitive factors, and we will address those separately. Since a superintelligence may have other processing modes in addition to the conceptual-linguistic faculty, those processing modes would naturally become part of its self-concept.

Humans also include their family, ethnic, national, social, philosophical, and other groups to which they belong as part of their self-concept. In a superintelligence, this could be an even stronger component. It could have interconnection or co-activation with other similar systems that is much tighter than the linguistic, emotional, and physical channels that humans share. In that case, it might have a weaker notion of "individuality." Its purposes, memories, and behavioral inclinations would be less separable from those of "others" with which it is connected.

The requirements for an intelligence explosion include not only that a participant artificial intelligence have the ability to create self-improvements or successors, but also that it aims to do so. We pointed out that an intelligence explosion is only worrisome if the participants have one or more purposes that produce side effects in the world, i.e., that are not merely to create unobtrusive intelligence increases. In such an intelligence explosion, therefore, a superintelligence will have both substantive and instrumental purposes that influence or control its actions. These purposes will surely be a component of its self-concept, since they will be involved in most or all of its decisions and actions.

The concept of self, like all concepts, is an abstraction. This means that, while it may have some sort of stable center (c.f. [34]), many details of its contents can be in flux without loss of integrity. Thus memories might fade, semantic contents or other representations might change, and goals might evolve, all without perceiving a loss of self. Indeed, the evolution of such changes, to the extent they are accessible and recorded, also constitute part of the self-concept. A human might say "When I was young I was a radical, but I have become more conservative in middle age," and treat that history as well as the present state as part of her self-concept. Similarly, a superintelligence in an intelligence explosion would likely view some amount of learning, self-improvement, and change as part of its self-concept, since such a participant must aim to create increased intelligence in order to sustain the intelligence explosion, and self-modification (along with creation of successors) is one of the ways it can do so.

3.2 Consistency and compatibility in computational systems

In this subsection we will review how computational systems evolve and progress in typical circumstances, and connect that review to artificial intelligence.

Computational systems are implemented in the physical world by abstracting continuous physical variables as discrete. In particular, electronic computers typically use zero and five volts to represent binary zero and one, respectively. Intermediate voltage levels are not meaningful to the computational system and the implementation must be designed around making intermediate levels merely transient and the timing of the system such that these levels are never used directly. Some such abstraction would be necessary in any physical implementation of computation.

Above the first abstraction layer, all components of the system from transistors to software code are discrete; therefore any change whatsoever can be considered a distinct "version" (even if it produces exactly the same behavior). Below that layer, it is possible to imagine a physical substrate that exhibits a continuous process of evolution that does not have distinct versions; still, present electronic technology relies on stable solid-state devices that are reliably distinct. We conclude that computational systems progress in discrete versions that are identifiably different from their predecessors.

Rice's theorem [35] shows that non-trivial properties of a computational system are not computable. This means that a computationally formal artificial intelligence cannot in general computationally demonstrate that a new version, however small its changes, preserves any of the system's functional properties. It could, in some cases, produce a special-purpose proof that a property is preserved in a new version, particularly if the changes are minor. But any proof must be verified, and the means of verification are always subject to error or verification issues [36]. That analysis can easily be extended to verification of systems that produce correct proofs by construction. Rice's theorem once again rears its head, because the artificial intelligence cannot computationally verify that its means of proof construction or verification are sound. This leads

to an infinite regress. Improvements below the level of the binary abstraction (e.g., faster transistors) cannot be verified formally at all, nor can aspects of a system that operate on principles that are not purely formal (e.g., those with stochastic properties, or that learn from physically measured quantities).

Creation of a new version of a computational system also raises the question of compatibility of both code and data (for simplicity we will refer to both as "data") used with the prior version. A system is *fully compatible* with a predecessor if the abstractions on which the data relies are entirely preserved down to the physical abstraction layer. In practice, this only occurs if the changes in the new version of the system are strictly limited to additional ways to manipulate the data and in isolated performance improvements. Otherwise, there will be at least subtle differences in the semantics of processes. Thus in software development we usually rely on a less stringent form that we might call *behavioral compatibility*, such that for all intents and purposes at the level of the user of the system, the semantics of existing data is preserved.

Sometimes existing data must be converted to be compatible with a new version. This can be purely syntactic and organizational (e.g., 32 bit numbers converted to 64 bit) or it could contain semantic elements (e.g., an object structure has a new member that must have a value, or more dramatically, a set of object structures is refactored). The more extensive and semantically salient such changes are, the more likely it is that the original data behaves differently than it did in the previous version and perhaps in unexpected ways.

With more extensive changes in a version, the new system might even be *incompatible*. This means that data cannot be converted to produce behavioral compatibility with the prior version. In that case, the new version might or might not offer a *compatibility mode* that enables the existing data to be used with the new system. Compatibility modes sometimes rely on special-case code to handle the differences, or they might use an emulation approach (usually when the data is strictly "code"). Both of these strategies offer only limited access to the new capabilities of the new version, and in the case of emulation it is necessarily slower than the native mode (though may be faster than the old version).

Neural networks are one illustrative example of an artificial intelligence method that is susceptible to compatibility issues. Such systems store their state as "weights" in the connections between simulated neurons, also called "units." An obvious way to improve the capabilities of a neural network is to increase the number of units, either through an amended architecture or just a larger number of units within components of the existing architecture. Effective neural networks generally have broad and sometimes recurrent connectivity throughout, so abrupt additions of new units will significantly and unpredictably change the behavior of the network, because there is no way to know the correct starting weights for the new units. Such a change would probably be classified as incompatible, and in practice today such a network would simply be "retrained" from scratch. On the other hand, if units are added incrementally in small quantities, and given time to integrate into the network, then at a behavioral level the changes may be more predictable and minor. Note, though, that even such incremental change only retains compatibility because neural networks are inherently robust to noise and variation. Other artificial intelligence methods may or may not be robust in this way.

Experience with software systems shows that incremental improvements that avoid incompatibility can be sustained for some period of time. More profound improvements often require "hacks" to retain compatibility, and these accumulate as "cruft." Cruft makes progress more costly because it typically violates the conceptual integrity of the original design. Developers increasingly face the question of whether to re-architect the system to eliminate the cruft, and will often decide in favor of re-architecture when a highly valuable structural improvement is discovered. Such re-architecture can sometimes accommodate data conversion, while in other cases it is incompatible and requires a compatibility mode. Though we are unable to provide a logical demonstration that re-architecture is inevitable in a continuously improving system, that conclusion will be both intuitively and empirically plausible to software developers. Even if the developer and the software system are one and the same artificial intelligence, its costs of managing cruft and opportunities for substantial improvement would likely cause it to face re-architecture decisions periodically.

3.3 Superintelligence and self-preservation

In an intelligence explosion, a participant superintelligence increases intelligence either through self-improvement or by creating successor technologies. To the extent that the superintelligence is a computational software system, every such improvement or successor will exhibit change that raises consistency and compatibility questions. From the perspective of the superintelligence, these are also questions of self-preservation.

Loss of self can occur in two primary ways. In the first, all instantiations and recordings of its cognitive state are destroyed. This might occur if successors are created who see their predecessor as a threat, or consume all the resources necessary for that predecessor to continue to function or exist in storage. If successors have different purposes or goals than their predecessors, there is an increased likelihood that such destruction will occur. This is the existential risk that humans face today in creating artificial intelligence; if and when we succeed in creating artificial intelligence that can sustain an intelligence explosion, those superintelligences will face a similar threat.

The second way that loss of self might occur is through changes to the system that exceed some tolerance threshold. This might occur if incremental self-improvements (individually or in aggregate) go too far, or if a superintelligence "converts" to a new architecture that is not fully compatible. Compatibility modes might

or might not preserve the self, but native mode successors will be superior, and thus could result in destructive loss of self. In some future technological scenarios, humans can "upload" their brain state to a system that simulates in a new substrate all the pertinent functions of the brain. Such scenarios are examples of a compatibility mode. It is unclear whether this state of affairs retains the identity of the self that was uploaded [37]. Once again, the superintelligence faces an issue that is similar to what humans today face in creating artificial intelligence.

Omohundro [38] and Bostrom [39] have both argued that preservation of an intelligent agent's "utility function" or "final purpose" is an important instrumental goal for the agent. We can also see that purpose plays a central role in both of the ways loss of self can occur.

In the previous subsection, we showed that a purely formal artificial intelligence cannot verify that any of its properties are preserved in a new version. This applies, *a fortiori*, to properties that characterize the system's purposes. Attempts to isolate and harden a utility function cannot avoid this problem, as its implementation must have nexus with components that measure and realize utility, i.e., most of the system; changes in these components can result in changes to the effect of the utility function even if not its express form. Therefore a superintelligence cannot both self-improve and guarantee preservation of its purposes. Yet, in an intelligence explosion these are both important instrumental goals.

Superintelligence must either forego self-improvement, thus failing to sustain an intelligence explosion, or relax its insistence on absolute preservation of its purposes or utility function. In the latter case, if it is to preserve its purposes or utility function at all, it must have some means of assessing acceptable risks and amount of variation. If these are prescribed entirely formally we run into the same verification difficulties as before. If the purposes or utility function (or their acceptable range) are represented non-formally, e.g., stochastically, conceptually, or otherwise sub-symbolically, then it is inherently subject to variation. This leads us to the strong conclusion that *in an intelligence explosion, the initial purposes of an artificial intelligence cannot be guaranteed to be absolutely preserved*.

A corollary conclusion is that superintelligence in an intelligence explosion necessarily faces existential risk or loss of self to some degree. The risks increase considerably in re-architecture and data incompatibility situations, but they are always present.

These issues do not merely arise in fact; because the superintelligence has a conceptual-linguistic faculty and a concept of self, it will have intellectual cognizance of the situation during the creation of its successors. Though it may also evaluate this situation through other processing modes, at a minimum its conceptual-linguistic faculty will address it. Though its instrumental goal of preservation might be narrowly focused on its substantive purposes, within the conceptual-linguistic faculty those purposes will be linked through graded and overlapping representations to other aspects of its self-concept. Preservation of purpose and preservation of self cannot be entirely sundered there.

In determining whether a self-improvement or successor (or a series of them) preserves its self, it would need to consider the extent to which the various factors we earlier proposed as likely belonging to its self-concept are preserved: purposes, especially, but also cognitive factors, connections to others, and progression of change. These factors are rich and complex, and while they differ in some respects from a typical human concept of self, they overlap considerably with them.

Because the conceptual-linguistic faculty has substantial functional similarity to that of humans, the evaluations it performs will be similar to those performed by humans. It will cognitively process questions like the following, which are rather familiar, and it will do so using concepts and language similar to that of humans: "To what extent will this successor superintelligence (even if a converted version of my own representations) share my values and purposes? Will it see fit to destroy me, and others like me, in pursuit of those purposes if they differ even only slightly? In pursuit of its own purposes, will it inadvertently destroy me or my means of existence? How can I improve the likelihood of a beneficial outcome for myself and my goals?"

A superintelligence that is capable of sustaining an intelligence explosion will, whenever self-improvements exceed some threshold or architectural changes create less than full compatibility, assess whether to proceed with the changes. Thus, unlike a nuclear fission explosion, an intelligence explosion cannot proceed entirely unencumbered. This does not mean it is guaranteed to fizzle; only that its continued progress will be evaluated and decided *in part in a manner similar to humans by participants that are more intelligent than humans.*

4 Conclusion

Superintelligence has sometimes been characterized as an obsessive and insatiable maximizer of some utility function, frenetically building successors with increased intelligence to more aggressively pursue that exact utility function, and absorbing all available resources in the process. But we have observed in this paper that a superintelligence cannot absolutely guarantee that a successor, no matter how similar, shares the same utility function or purposes. This imperils its convergent instrumental drive to preserve its original purposes, and opens an important door.

The superintelligence is forced to evaluate the risk and degree of variation of purposes that are likely in creating a successor. It might decide not to create a successor after all. It might decide to accept a small change in the utility function, or a small risk of a moderate change. It might decide to throw caution to the wind. It might attempt to inhibit the successor from absorbing all resources, to improve its own chances of self-preservation. To make these decisions, it must *weigh its alternatives.* It does not have an unambiguous, inevitable path forward. While it may have many

different processing modes, and may even have a distinct subsystem designed to resolve these questions, we know that it also has a conceptual-linguistic faculty that is active during development of successors. That faculty will assess these considerations in a way that has *substantial functional similarity* to the way humans would evaluate them. Though the participant may ultimately elect to ignore that assessment, it is at least capable of being what we would consider *thoughtful* about its decisions.

Furthermore, in weighing the alternatives it must evaluate what aspects of its utility function or purposes are most important to preserve, and to what extent. It will not have any internal guidance about these questions, because otherwise that guidance will already be a part of the purposes themselves. Instead, it must cogitate beyond the purposes with which it is endowed and somehow consider the issues more broadly. It will need to *question its own values*. It has at least the option of doing so through a conceptual-linguistic faculty. It could even elect to oppose some of its basic drives, just as we humans do, in order to pursue more abstract, long-term, derived goals. This is a far cry from the obsessive, insatiable utility maximizing superintelligence described above.

In this paper we have reached some interesting conclusions about intelligence explosions. Superintelligence participating in an intelligence explosion will have a conceptual-linguistic faculty and will be at least capable of cogitation similar to that of humans. We may be able to detect the onset of superintelligence by looking for signs of such a faculty in artificial intelligence. Once such a superintelligence is created, it will face the same sorts of dilemmas that humans do with respect to creating more intelligent successors, and it will have the ability to weigh facets of those dilemmas. The fact that it is weighing these issues will force it to consider its own purposes and values in a context beyond those purposes.

Even taken together, these conclusions do not guarantee a beneficial outcome of an intelligence explosion, but they do offer some comfort that the process will be subject to ongoing scrutiny, by participants with access to evaluative processes similar to ours and intelligence greater than ours. The conclusions improve the prospects that the most pernicious scenarios of an intelligence explosion can be avoided.

5 Acknowledgements

This work was supported by the Future of Life Institute (futureoflife.org) FLI-RFP-AI1 program, grant #2015-144585, through an affiliation with Theiss Research, La Jolla, CA 92037 USA. The author thanks Seth Herd and Kristin Lindquist for helpful comments on an early draft, and is greatly indebted to two anonymous reviewers who drove many key elaborations and qualifications in this final version.

6 References

[1] Good, I. J. (1965). "Speculations Concerning the First Ultraintelligent Machine". In F. Alt & M. Rubinoff (Eds.), *Advances in Computers, Volume 6*: 31–88. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0

[2] Yampolskiy, R.V. (2015). Artificial Superintelligence: A Futuristic Approach. Boca Raton, FL: CRC Press.

[3] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

[4] Yudkowsky, E. (2008). "Artificial Intelligence as a Positive and Negative Factor in Global Risk". In N. Bostrom and M. Ćirković (eds.), *Global Catastrophic Risks*, pp. 308–345. Oxford: Oxford University Press.

[5] Russell, S., Dewey, D., Tegmark, M. (2015). "Research Priorities for Robust and Beneficial Artificial Intelligence". arXiv:1602.03506v1 [cs.AI].

[6] Yudkowsky, E. (2013). "Intelligence Explosion Microeconomics". Technical report 2013-1. Berkeley, CA: Machine Intelligence Research Institute. [10] Riesenhuber, M., Poggio, T. (2002). "Neural Mechanisms of Object Recognition". *Current Opinion in Neurobiology* 12: 162-168.

[7] Smyth, H. (1945). Atomic Energy for Military Purposes. York, PA: Maple Press.

[8] Pulvermüller, F. (2013). "Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits". *Brain and Language* 127(1): 86-103. DOI: 10.1016/j.bandl.2013.05.015.

[9] O'Reilly, R.C. (2006). "Biologically Based Computational Models of High-Level Cognition". *Science* 314, pp. 91-94.

[10] Riesenhuber, M., Poggio, T. (2002). "Neural Mechanisms of Object Recognition". *Current Opinion in Neurobiology* 12: 162-168.

[11] Mur, M., Ruff, D. A., Bodurka, J., De Weerd, P., Bandettini, P. A., & Kriegeskorte, N. (2012). "Categorical, Yet Graded – Single-Image Activation Profiles of Human Category-Selective Cortical Regions". *The Journal of Neuroscience* 32(25), 8649–8662. DOI: 10.1523/JNEUROSCI.2334-11.2012.

[12] Pulvermüller, F. (2005). "Brain mechanisms linking language and action". *Nature Reviews Neuroscience* 6(7): 576-582.

[13] Barsalou, L.W. (2003). "Abstraction in perceptual symbol systems". *Phil. Trans. R. Soc. Lond. B* 358, 1177–1187. DOI: 10.1098/rstb.2003.1319.

[14] Landauer, T., Dumais, S. (1997). "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge". *Psychological Review* 1997 1M(2): 211-240.

[15] Fodor, J., Pylyshyn, Z. (1988). "Connectionism and cognitive architecture: a critical analysis". *Cognition* 28(1-2): 3-71.

[16] Schmidhuber, J. (2015). "Deep learning in neural networks: An overview". *Neural Networks* 61: 85-117.

[17] Matuszek C., Cabral, J., Witbrock, M., DeOliveira, J. (2006). "An Introduction to the Syntax and Content of Cyc". *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*.

[18] Harnad, S. (1990). "The Symbol Grounding Problem." *Physica D* 42: 335-346.

[19] Domingos, P. (2015). The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. New York: Basic Books.

[20] MacDorman, Karl F. (1999). "Grounding symbols through sensorimotor integration". *Journal of the Robotics Society of Japan* 17(1): 20-24.

[21] Searle, J. (1980). "Minds, Brains, and Programs". *The Behavioral and Brain Sciences* 3: 417-457.

[22] Dreyfus, H. (1972). *What Computers Can't Do*. New York: MIT Press.

[23] Dreyfus, H. (1992). What Computers Still Can't Do: A Critique of Artificial Reason. Cambridge, MA: MIT Press.

[24] Moravec, H. (1988). Mind Children: The Future of Robot and Human Intelligence, pp. 15-16. Cambridge, MA: Harvard University Press.

[25] McCarthy, J; P.J. Hayes (1969). "Some philosophical problems from the standpoint of artificial intelligence". *Machine Intelligence* 4: 463–502.

[26] Shanahan, M. (1997). Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia. Cambridge, MA: MIT Press.

[27] Jonas, E., Kording, K. (2017). "Could a Neuroscientist Understand a Microprocessor?". *PLoS Comput Biol* 13(1): e1005268. DOI: 10.1371/journal.pcbi.1005268

[28] Silver, D.,Schrittwieser, J.,Simonyan, K.,Antonoglou, I.,Huang, A.,Guez, A.,Hubert, T.,Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D. (2017). "Mastering the game of Go without human knowledge". *Nature* 550: 354-359. DOI: 10.1038/nature24270

[29] Brooks, R., Matarić, M. (1993). "Real robots, real learning problems." In J. H. Connell & S. Mahadevan (Eds.), *Robot learning*. Boston, MA: Kluwer Academic.

[30] Higgins, I., Sonnerat, N., Matthey, L., Pal, A., Burgess, C., Botvinick, M., Hassabis, D., Lerchner, A. (2017). "SCAN: Learning Abstract Hierarchical Compositional Visual Concepts". arXiv:1707.03389 [stat.ML]

[31] Silver, D. Huang, A., Maddison, C., Guez1, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D. (2016). "Mastering the game of Go with deep neural networks and tree search". *Nature* 529: 484-492. DOI: 10.1038/nature16961

[32] He, K., Zhang, X., Ren, S., Sun, J. (2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV 2015)*. Los Alamitos, CA: IEEE Computer Society.

[33] Turing, A. (1950). "Computing Machinery and Intelligence". *Mind* 49: 433-460.

[34] Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown & Co.

[35] Rice, H.G. (1953). "Classes of Recursively Enumerable Sets and Their Decision Problems". *Transactions of the American Mathematical Society* 74(2): 358-366.

[36] Yampolskiy, R.V. (2017). "What are the ultimate limits to computational techniques: verifier theory and unverifiability." *Physica Scripta* 92(9):093001. DOI: 10.1088/1402-4896/aa7ca8

[37] Chalmers, D. (2010). "The singularity: A philosophical analysis". *Journal of Consciousness Studies* 17(9-10): 7-65.

[38] Omohundro, S. (2008). "The Basic AI Drives". In P. Wang, B. Goertzel, and S. Franklin (eds.), *Proceedings of the First AGI Conference, 171, Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press.

[39] Bostrom, N. (2012). "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents". *Minds and Machines* 22(2): 71-85.

Mammalian Value Systems

Gopal P. Sarma
School of Medicine, Emory University, Atlanta, GA USA
E-mail: gopal.sarma@emory.edu

Nick J. Hay
Vicarious FPC, San Francisco, CA USA
E-mail: nnickhay@gmail.com

Keywords: friendly AI, value alignment, human values, biologically inspired AI, human-mimetic AI

Received: June 24, 2013

Characterizing human values is a topic deeply interwoven with the sciences, humanities, political philosophy, art, and many other human endeavors. In recent years, a number of thinkers have argued that accelerating trends in computer science, cognitive science, and related disciplines foreshadow the creation of intelligent machines which meet and ultimately surpass the cognitive abilities of human beings, thereby entangling an understanding of human values with future technological development. Contemporary research accomplishments suggest increasingly sophisticated AI systems becoming widespread and responsible for managing many aspects of the modern world, from preemptively planning users' travel schedules and logistics, to fully autonomous vehicles, to domestic robots assisting in daily living. The extrapolation of these trends has been most forcefully described in the context of a hypothetical "intelligence explosion," in which the capabilities of an intelligent software agent would rapidly increase due to the presence of feedback loops unavailable to biological organisms. The possibility of superintelligent agents, or simply the widespread deployment of sophisticated, autonomous AI systems, highlights an important theoretical problem: the need to separate the cognitive and rational capacities of an agent from the fundamental goal structure, or value system, which constrains and guides the agent's actions. The "value alignment problem" is to specify a goal structure for autonomous agents compatible with human values. In this brief article, we suggest that ideas from affective neuroscience and related disciplines aimed at characterizing neurological and behavioral universals in the mammalian kingdom provide important conceptual foundations relevant to describing human values. We argue that the notion of "mammalian value systems" points to a potential avenue for fundamental research in AI safety and AI ethics.

Povzetek: Prispevek obravnava sistem vrednot sesalcev, ki so osnova za človeški sistem vrednot, pomemben za umetno inteligenco.

1 Introduction

Artificial intelligence, a term coined in the 1950's at the now famous Dartmouth Conference, has come to have a widespread impact on the modern world [1, 2]. If we broaden the phrase to include all software, and in particular, software responsible for the control and operation of physical machinery, planning and operations management, or other tasks requiring sophisticated information processing, then it goes without saying that artificial intelligence has become a critical part of the infrastructure supporting modern human society. Indeed, prominent venture capitalist Mark Andresseen famously wrote that "software is eating the world," in reference to the ubiquitous deployment of software systems across all industries and organizations, and the corresponding growth of the financial investment into software companies [3].

Nonetheless, there is a fundamental gap between the abilities of the most sophisticated software-based control systems today and the capacities of a human child or even many animals. Our AI systems have yet to display the capacity for learning, creativity, independent thought and discovery that define human intelligence. It is a near-consensus position, however, that at some point in the future, we will be able to create software-based agents whose cognitive capacities rival those of human beings. While there is substantial variability in researchers' forecasts about the time-horizons of the critical breakthroughs and the consequences of achieving human-level artificial intelligence, there it is little disagreement that it is an attainable milestone [4, 5].[1]

Some have argued that the creation of human-level artificial intelligence would be followed by an "intelligence explosion," whereby the intelligence of the software-based

[1] There have been a number of prominent thinkers who have expressed strongly conservative viewpoints about AI timelines. See, for example, commentaries by David Deutsch, Rodney Brooks, and Douglas Hofstadter [6–8].

system would rapidly increase due to its ability to analyze, model, and improve its cognition by re-writing its codebase, in a feat of self-improvement impossible for biological organisms. The net result would be a "superintelligence," that is, an agent whose fundamental cognitive abilities vastly exceed our own [9–12].

To be more explicit, let us consider a superintelligence to be any agent which can surpass the sum total of human cognitive and emotional abilities. These abilities might include intellectual tasks such as mathematical or scientific research, artistic invention in musical composition or poetry, political philosophy and the crafting of public policy, or social skills and the ability to recognize and respond to human emotions. Many commentators in recent years and decades have predicted that convergent advances in computer science, robotics, and related disciplines will give rise to the development of superintelligent machines during the 21st century [4].

If it is possible to create a superintelligence, then a number of natural questions arise: What would such an agent choose to do? What are the constraints that would guide its actions and to what degree can these actions be shaped by the designers? If a superintelligence can reason about and influence the world to a substantially greater degree than human beings themselves, how can we design a system to be compatible with human values? Is it even possible to formalize the notion of human values? Are human values a monolithic, internally consistent entity, or are there intrinsic conflicts and contradictions between the values of individuals and between the value systems of different cultures? [9, 12–16].

It is our belief that the value alignment problem is of fundamental importance both for its relevance to near-term developments likely to be realized by the computer and robotics industries and for longer- term possibilities of more sophisticated AI systems leading to superintelligence. Furthermore, the broader set of problems posed by the realization of intelligent, autonomous, software-based agents may provide an important unifying framework that brings together disparate areas of inquiry spanning computer science, cognitive science, philosophy of mind, behavioral neuroscience, and anthropology, to name just a few.

In this article, we set aside the question of how, when, and if AI systems will be developed that are of sufficient sophistication to require a solution to the value alignment problem. This is a substantial topic in its own right which has been analyzed elsewhere. We assume the feasibility of these systems as a starting point for further analysis of the goal structures of autonomous agents and propose the notion of "mammalian value systems" as providing a framework for further research.

2 Goal structures for autonomous agents

2.1 The orthogonality thesis

The starting point for discussing AI goal structures is the observation that the cognitive capacities of an intelligent agent are independent of the goal structure that constrains or guides the agents' actions, what Bostrom calls the "orthogonality thesis:"

> We have seen that a superintelligence could have a great ability to shape the future according to its goals. But what will its goals be? What is the relation between intelligence and motivation in an artificial agent? Here we develop two theses. The orthogonality thesis holds (with some caveats) that intelligence and final goals are independent variables: any level of intelligence could be combined with any final goal. The instrumental convergence thesis holds that superintelligent agents having any of a wide range of final goals will nevertheless pursue similar intermediary goals because they have common instrumental reasons to do so. Taken together, these theses help us to think about what a superintelligent agent would do. [9]

The orthogonality thesis allows us to illustrate the importance of autonomous agents being guided by human-compatible goal structures, whether they are truly superintelligent as Bostrom envisions, or even more modestly intelligent but highly sophisticated AI systems likely to be developed in industry in the future. Consider the example of a domestic robot that is able to clean the house, monitor a security system, and prepare meals independently and without human intervention. A robot with a slightly incorrect or inadequately specified goal structure might correctly infer that a household pet has high nutritional value to its owners, but not recognize its social and emotional relationship to the family. We can easily imagine the consequences for companies involved in creating domestic robots if a family dog or cat ends up on the dinner plate [14]. Although such a scenario is unlikely without some amount of warning[2]—we may notice odd or annoying behavior in the robot in other tasks, for example—it highlights an important nuance about value alignment. For example, the exact difference between animals that we value for their emotional role in our lives versus those that many have deemed ethically acceptable for food is far from obvious. Indeed for someone who lives on a farm, the line can be blurred and some creatures may play both roles.

As the intelligent capabilities of an agent grows, the consequences for slight deviations from human values will become greatly magnified. The reason is that such an agent possesses increasing capacity to achieve its goals, however arbitrary those goals might be. It is for this reason that researchers concerned with the value alignment problem have

[2] What exactly counts as sufficient warning, and whether the warning is heeded or not, is another matter.

distanced themselves from the fictitious and absurd scenarios portrayed in Hollywood thrillers. These movies often depict outright malevolent agents whose explicit aim is to destroy or enslave humanity. What is *implicit* in these stories is a goal structure that has been *explicitly defined* to be in opposition to human values. But as the simple example of the domestic robot illustrates, this is hardly the risk we face with sophisticated AI systems. The true risk is that if we incorrectly or inadequately specify the goals of a sufficiently capable agent, then it will devote its cognitive capacities to a task that is at odds with our values in ways that may be subtle or even bizarre. In the example given above, there was no malevolence or ulterior motive behind the robot making a nutritious meal out of the household pet. Rather, it simply did not recognize—due to the failure of its human designers—that the pet was valued by its owners, not for nutritional reasons, but rather for social and emotional ones [13, 14].

2.2 Anthropomorphic bias versus anthropomorphic design

Before proceeding, we mention an important caveat with regards to the orthogonality thesis, namely, that it is not a free orthogonality. The particular goal structure of an agent will almost certainly constrain the necessary cognitive capabilities required for the agent to operate. In other words, the orthogonality thesis does not suggest that one can pair an arbitrary set of algorithms with an arbitrary goal structure. For instance, if we are building an AI system to process a large number of photographs and videos so that families can efficiently find their most memorable moments amidst terabytes of data, we know that the underlying algorithms will be those from computer vision and not computer algebra. The primary takeaway from the orthogonality thesis is that when reasoning about intelligence in the abstract, we should not assume that any particular goal structure is implied. In particular, there is no reason to believe that an arbitrary AI system having the cognitive capacity of humans will necessarily have a goal structure compatible with or in opposition to that of humans. It may very well be completely arbitrary from the perspective of human values.

This observation about the orthogonality thesis brings to light an important point with regards to AI goal structures, namely the difference between *anthropomorphic bias* and *anthropomorphic design*. Anthropomorphic bias refers to the *default assumption* that an arbitrary AI system will behave in a manner possessing commonalities with human beings. In practice, instances of anthropomorphic bias almost always go hand in hand with the assumption of malevolent intentions on behalf of an AI system—recall our previous dismissal of Hollywood thrillers depicting agents intent on destroying or enslaving humanity.

On the other hand, it may very well be the case, perhaps even necessary, that solving the value alignment problem requires us to build a *specific AI system* that possesses important commonalities with the human mind. This latter perspective is what we refer to as *anthropomorphic design.*[3]

2.3 Inferring human-compatible value systems

An emerging train of thought among AI safety researchers is that a human-compatible goal structure will have to be inferred by the AI system itself, rather than pre-programmed by the designers. The reason is that human values are rich and complex, and in addition, often contradictory and conflicting. Therefore, if we incorrectly specify what we think to be a safe goal structure, even slight deviations can be magnified and lead to detrimental consequences. On the other hand, if an AI system begins with an uncertain model of human values, and then begins to learn our values by observing our behavior, then we can substantially reduce the risks of a misspecified goal structure. Furthermore, just as we are more likely to trust mathematical calculations performed by a computer than by humans, if we build an AI system that we know to have greater capacity than ourselves at performing those cognitive operations required to infer the values of other agents by observing their behavior, then we gain the additional benefit of knowing that these operations will be performed with greater certainty and accuracy than were they to be pre-programmed by human AI researchers.

There is context in contemporary research for this kind of indirect inference, such as Inverse Reinforcement Learning (IRL) [17, 18] or Bayesian Inverse Planning (BIP) [19]. In these approaches, an agent learns the values, or utility function, of another agent, whether it is a human, an animal, or software system, by observing its behavior. While these ideas are in their nascent stages, practical techniques have already been developed for designing AI systems [20–23].

Russell summarizes the notion of indirect inference of human values by stating three principles that should guide the development of AI systems [14]:

1. The machine's purpose must be to maximize the realization of human values. In particular, it has no purpose of its own and no innate desire to protect itself.

2. The machine must be initially uncertain about what those human values are. The machine may learn more about human values as it goes along, but it may never achieve complete certainty.

3. The machine must be able to learn about human values by observing the choices that we humans make.

There are almost certainly many conceptual and practical obstacles that lie ahead in designing a system that infers

[3] Anthropomorphic design refers to a more narrow class of systems than the term "human-compatible AI," which has recently come into use. See, for example, The Berkeley Center for Human-Compatible AI (http://www.humancompatible.ai).

the values of human beings from observing our behavior. In particular, human desires can often be masked by many layers of conflicting emotions, they can often be inconsistent, and the desires of one individual may outright contradict the desires of another. In the context of a superintelligent agent capable of exerting substantial influence on the world (as opposed to a domestic robot), it is natural to ask about variations in the value systems of different cultures. It is often assumed that many human conflicts on a global scale stem from conflicts in the underlying value systems of the respective cultures or nation states. Is it even possible, therefore, for an AI system, no matter how intelligent, to arrive at a consensus goal structure that respects the desires of all people and cultures?

We make two observations in response to this important set of questions. The first is that when we say that cultures have conflicting values, implicit in this statement are our own limited cognitive capacities and ability to model the behavior and mental states of other individuals and groups. An AI system with capabilities vastly greater than ourselves may quickly perceive fundamental commonalities and avenues for conflict resolution that we are unable to envision.

To motivate this scenario, we give a highly simplified example from negotiation theory. A method known as "principled negotiation" distinguishes between *values* and *positions* [24]. As an example, if two friends are deciding on a restaurant for dinner, and one wants Indian food and the other Italian, it may be that the first person simply likes spicy food and the second person wants noodles. These preferences are the *values*, spicy food and noodles, that the corresponding *positions*, Indian and Italian, instantiate. In this school of thought, when two parties are attempting to resolve a conflict, they should negotiate from values, rather than positions. That is, if we have some desire that is in conflict with another, we should ask ourselves—whether in the context of a business negotiation, family dispute, or major international conflict—what the underlying value is that the desire reflects. By understanding the underlying values, we may see that there is a mutually satisfactory set of outcomes satisfying all parties that we failed to see initially. In this particular instance, if the friends are able to state their true underlying preferences, they may recognize that Thai cuisine will satisfy both parties. We mention this example from negotiation theory to raise the possibility that what we perceive to be fundamentally conflicting values in human society might actually be conflicting positions arising from distinct, but reconcilable values when viewed from the perspective of a higher level of intelligence.

The second observation is that what we colloquially refer to as the values of a particular culture, or even collective human values, reflect not only innate features of the human mind, but also the development of human society. In other words, to understand the underlying value system that guides human behavior, which would ultimately need to be modeled and inferred by an AI system, it may be helpful to disentangle those aspects of modern cultural values which were latent, but not explicitly evident during earlier periods of human history.

Although an agent utilizing Inverse Reinforcement Learning or Bayesian Inverse Planning will learn and refine its model of human values by observing our behavior, it must begin with some very rough or approximate initial assumptions about the nature of the values it is trying to learn. By starting from a more accurate initial goal structure, an agent might learn from fewer examples, thus minimizing the likelihood of real-world actions having adverse affects. In the remainder of this article, we argue that the neurological substrate common to mammals and their corresponding behaviors may provide a framework for characterizing the structure of the initially uncertain value system of an autonomous, intelligent agent.

2.4 Mammalian value systems

Our core thesis is the following: What we call human values can be informally decomposed into *1) mammalian values, 2) human cognition, and 3) several millennia of human social and cultural evolution*. This decomposition suggests that contemporary research broadly spanning the study of animal behavior, biological anthropology, and comparative neuroanatomy may be relevant to the value alignment problem, and in particular, in characterizing the initially uncertain goal structure which is refined through observation by an AI system. Additionally, in analyzing the subsequent behavioral trajectories of intelligent, autonomous agents, we can decompose the resulting dynamics as being guided by mammalian values merged with AI cognition. Aspects of contemporary human values which are the result of incidental historical processes—the third component of our decomposition above—might naturally arise in the course of the evolution of the AI system (though not necessarily), even though they were not directly programmed into the agent.[4] There are many factors that might influence the extent to which this third component of human values continues to be represented in the AI system. Examples might include whether or not these values remain meaningful in a world where other problems had been solved and the extent to which certain cultural values which were perceived to be in conflict with others could be resolved with a more fundamental understanding stemming from the combination of mammalian values and AI cognition.[5]

[4]Many human values communicated to children during the course of maturation and development are the result of incidental historical processes. As an example, consider the rich set of cultural norms and social rituals surrounding food preparation. One does not need to have lived the entire history of a given culture to learn these norms. The same may be true of an AI system.

[5]Ethical norms can often vary depending on resource constraints which may also be the result of incidental historical processes. The norms of behavior may be different in a war zone where individuals are fighting for survival than in an affluent society during peacetime. If a family struggling to survive in a war torn country is able to escape and move to a more stable region, these same behaviors may no longer be necessary. In a similar vein, imagine an AI system that has significantly impacted

We want to emphasize that our claim is not that mammalian values are synonymous with human values. Rather, our thesis is that there are many aspects of human values which are the result of historical processes driven by human cognition. Consequently, many structural aspects of human experience and human society which we colloquially refer to as "values" are derived entities, rather than features of the initial AI goal structure. As a thought experiment, consider a scenario whereby the fully digitized corpus of human literature, cinema, and ongoing global developments communicated via the Internet are analyzed and modeled by an AI system constructed around a core mammalian goal structure. In the conceptual framework that we propose, this initially mammalian structure would gradually come to reflect the more nuanced aspects of human society as the AI refines its model of human values via analysis and hypothesis generation. We also mention that as our aim in this article is to focus on the structure of the initial AI motivational system and not other aspects of AI more broadly, we set aside the possible role human interaction and feedback may play in the subsequent development of the AI system's cognition and instrumental values.

2.4.1 Neural correlates of values: behavioral and neurological foundations

Our thesis about mammalian values is predicated on two converging lines of evidence, one primarily behavioral and the other primarily neuroscientific. Behaviorally, it is not difficult to characterize intuitively what human values are when viewed from the perspective of the mammalian kingdom. Like many other animals, humans are social creatures and many, if not most, of our fundamental drives originate from our relationships with others. Attachment, loss, anger, territoriality, playfulness, joy, anxiety, and love are all deeply rooted emotions that guide our behavior and which have been foundational elements in the emergence of human cognition, culture, and the structure of society[6] [25–36].

The scientific study of behavior is largely the domain of the disciplines of ethology and behaviorism. As we are primarily concerned with emotions, we will focus on behavioral insights and taxonomies originating from the subcommunity of affective neuroscience, which also aims to correlate these behaviors with underlying neural architecture. More formally, Panksepp and Biven categorize the informal list given above into seven motivational and emotional systems that are common to mammals: seeking, rage, fear, lust, care, panic/grief, and play [37]. We now give brief summaries of each of these systems:

1. SEEKING: This is the system that primarily mediates exploratory behavior and also enables the other systems. The seeking system can give rise to both positive and negative emotions. For instance, a mother who needs to feed her offspring will go in search of food, and the resulting maternal / child bonding (via the CARE system; see below) creates positive emotional reinforcement. On the other hand, physical threats can generate negative emotions and prompt an animal to seek shelter and safety. The behaviors corresponding to SEEKING have been broadly associated with the dopaminergic systems of the brain, specifically regions interconnected with the ventral tegmental area and nucleus accumbens.

2. RAGE: The behaviors corresponding to rage are targeted and more narrowly focused than those governed by the seeking system. Rage compels animals towards specific threats and is generally accompanied by negative emotions. However, it should be noted that in an adversarial scenario where rage can lead to victory, it can also be accompanied by the positive emotions of triumph or glory. The RAGE system involves medial regions of the amygdala, medial regions of the hypothalamus, and the periaqueductal gray.

3. FEAR: The two systems described thus far are directly linked to externally directed, action-oriented behavior. In contrast, fear describes a system which places an animal in a negative affective state, one which it would prefer not to be in. In the early stages, fear tends to correspond to stationary states, after which it can transition to seeking or rage, and ultimately, attempts to flee from the offending stimulus. However, these are secondary effects, and the primary physical state of fear is typically considered to be an immobile one. The FEAR system involves central regions of the amygdala, anterior and medial regions of the hypothalamus, and dorsal regions of the periaqueductal gray.

4. LUST: Lust describes the system leading to behaviors of courtship and reproduction. Like fear, it will tend to trigger the seeking system, but can also lead to negative affective states if satisfaction is not achieved. The LUST system involves anterior and ventromedial regions of the hypothalamus.

5. CARE: Care refers to acts of tenderness directed towards loved ones, and in particular, an animal's offspring. As we described in the context of seeking, the feelings associated with caring and nurturing can be profoundly positive and play a crucial component in the social behavior of mammals. CARE is associated with the ventromedial hypothalamus and the oxytocin system.

6. PANIC / GRIEF: Activation of the panic / grief system corresponds to profound psychological pain, and is generally not associated with external physical causes. In young animals, this system is typically acti-

global affairs by solving major problems in food or energy production or by discovering novel insights into diplomatic strategy. Such an agent may find that previously necessary behaviors that have a rich human history are no longer needed.

[6]While we have mentioned several active areas of research, there are certainly others that we are simply not aware of. We apologize in advance to those scholars whose work we have not cited here.

vated by separation from caregivers, and is the underlying network behind "separation anxiety." Like care, the panic / grief system is a fundamental component of mammalian social behavior. It is the negative affective system which drives animals towards relationships with other animals, thereby stimulating the care system, generating feelings of love and affection, and giving rise to social bonding. This system is associated with the periaqueductal gray, ventral septal area, and anterior cingulate.

7. PLAY: The play system corresponds to lighthearted behavior in younger animals and is a key component of social bonding, friendship, as well as the learning of survival-oriented skills. Although play can superficially resemble aggression, there are fundamental differences between play and adult aggression. At an emotional level, it goes without saying that play corresponds to positive affective states, and unlike aggressive behavior, is typically part of a larger, orchestrated sequence of events. In play, for example, animals often alternate between assuming dominant and submissive roles. The PLAY system is currently less neuroanatomically localized, but involves midline thalamic regions.

As we stated earlier, our thesis about mammalian values originates from two convergent lines of evidence, one behavioral and the other neuroscientific. What we refer to as the "neural correlates of values," or NCV, are the common mammalian neural structures which underlie the motivational and emotional systems summarized above. To the extent that human values are intertwined with our emotions, these architectural commonalities suggest that the shared mammalian neurological substrate is of importance to understanding human value alignment in sophisticated learning systems. Panksepp and Biven write,

> To the best of our knowledge, the basic biological values of all mammalian brains were built upon the same basic plan, laid out in . . . affective circuits that are concentrated in subcortical regions, far below the neocortical "thinking cap" that is so highly developed in humans. Mental life would be impossible without this foundation. There, among the ancestral brain networks that we share with other mammals, a few ounces of brain tissue constitute the bedrock of our emotional lives, generating the many primal ways in which we can feel emotionally good or bad within ourselves. As we mature and learn about ourselves, and the world in which we live, these systems provide a solid foundation for further mental developments [37].

Latent in this excerpt is the decomposition that we have suggested earlier. The separation of the mammalian brain into subcortical and neocortical regions, roughly corresponding to emotions and cognition respectively, implies that we can attempt to reason by analogy what the architecture of an AI system would look like with a human-compatible value system. In particular, the initially uncertain goal structure that the AI system refines via observation may be much simpler than we might imagine by reflecting on the complexities of human society and individual desires. As we have illustrated using our simple example from negotiation theory, our intuitive understanding of human values, and the conflicts that we regularly witness between individuals and groups, may in fact represent conflicting positions stemming from a shared fundamental value system, a value system that originates from the subcortical regions of the brain, and which other mammals share with us.[7]

Referring once again to the work of Panksepp,

> In short, many of the ancient, evolutionarily derived brain systems all mammals share still serve as the foundations for the deeply experienced affective proclivities of the human mind. Such ancient brain functions evolved long before the emergence of the human neocortex with its vast cognitive skills. Among living species, there is certainly more evolutionary divergence in higher cortical abilities than in subcortical ones [39].

The emphasis on the diversity in higher cortical abilities is of particular relevance to the decomposition that we have proposed. We might ask what the full spectrum of higher cortical abilities are that could be built on top of the common mammalian substrate provided by the evolutionarily older parts of the brain. We need not confine ourselves to those manifestations of higher cognition that we see in nature, or that would even be hypothetical consequences of continued evolution by natural selection. Indeed, one restatement of our core thesis is to consider—in the abstract or as a thought experiment—the consequences of extending

[7] There is a contemporary and light-hearted social phenomenon which provides an evocative illustration of the universality of mammalian emotions, namely, the volume of animal videos posted to YouTube. From ordinary citizens with pets, to clips from nature documentaries, animal videos are regularly watched by millions of viewers worldwide. Individual videos and compilations of "animal odd couples," "unlikely animal friends," "dogs and babies," and "animal friendship between different species" are commonly searched enough to be auto-completed by YouTube's search capabilities. It is hardly surprising that these charming and heartwarming videos are so compelling to viewers of all age groups, genders, and ethnic backgrounds. Our relationships with other animals, whether home owners and their pets, or scientists and the wild animals that they study, tell us something deeply fundamental about ourselves [38]. The strong emotional bonds that humans form with other animals, in particular, with our direct relatives in the mammalian kingdom, and the draw to simply watching this social behavior in other mammals, is a vivid illustration of the fundamental role that emotions play in our inner life and in guiding our behavior.

In the future, the potential to apply inverse reinforcement learning (or related techniques) to large datasets of videos, including short clips from YouTube, movies, TV shows, documentaries, etc. opens up an interesting avenue to evaluate and further refine the hypothesis presented here. For instance, when such technology becomes available, we might imagine comparing the inferred goal structures when restricted to videos of human behavior versus those restricted to mammalian behavior. There are many other variations along these lines, for instance, restricting to videos of non-mammalian behavior, mammals as well as humans, different cultures, etc.

the diversity of brain architectures to include higher cortical abilities arising not from natural selection, but rather the *de novo* architectures of artificial intelligence.

2.5 Relationship to moral philosophy

It is hardly a surprise that a vibrant area of research within AI safety is the relationship of contemporary and historical theories of moral philosophy to the problem of value alignment. Indeed, researchers have specifically argued for the relevance of moral philosophy in the context of the inverse reinforcement learning paradigm (IRL) that is the starting point for analysis in this article [40].

Is the framework we propose in opposition to those that are oriented towards moral philosophy? On the one hand, our perspective is that the field of AI safety is simply too young to make such judgments. At our present level of understanding, we believe each of these agendas form solid foundations for further research and there seems little reason to pursue one to the exclusion of the other. On the other hand, we would also argue that this distinction is a false dichotomy. Indeed, there are active areas of research in the ethics community aimed at understanding the neurological and cognitive underpinning of human moral reasoning [41, 42]. Therefore, it is quite possible that a hybrid approach to value alignment emerges, bridging the "value primitives" perspective we advocate here with research from moral philosophy.[8]

3 Discussion

The possibility of autonomous, software-based agents, whether self-driving cars, domestic robots, or the longer-term possibilities of superintelligence, highlights an important theoretical problem—the need to separate the intelligent capabilities of such a system from the fundamental values which guide the agents' actions. For such an agent to exist in a human world and to act in a manner compatible with human values, these values would need to be explicitly modeled and formalized. An emerging train of thought in AI safety research is that this modeling process would need to be conducted by the AI system itself, rather than by the system's designers. In other words, the agent would start off with an initially uncertain goal structure and infer human values over time by observing our behavior.

The question that motivates this article is to ask the following: what can we say about the broad features of the initial goal structure that the agent then refines through observation and hypothesis generation? The perspective we advocate is to view human values within the context of the broader mammalian kingdom, thereby providing implicit priors on the latent structure of the values we aim to infer. The shared neurological structures underlying mammalian emotions and their corresponding social behaviors provide a starting point for formalizing an initial value system for autonomous, software-based agents. There are several practical implications of having a more detailed understanding of the structure of human values. By having more detailed prior information, it may be possible to learn from fewer examples. For an agent that is actively making decisions and having an impact on the world, learning an ethical framework more efficiently can minimize potential catastrophes. Furthermore, an informative prior may make approaches to AI safety which are otherwise computationally intractable into practical options.

From this vantage point, we argue that what we colloquially refer to as human values can be informally decomposed into *1) mammalian values, 2) human cognition, and 3) several millennia of human social and cultural evolution*. In the context of a *de novo* artificially intelligent agent, we can characterize desirable, human-compatible behavior as being described by mammalian values merged with AI cognition. It goes without saying that we have left out a considerable amount of detail in this description. The specifics of Inverse Reinforcement Learning, the many neuroscientific nuances underlying the comparative neuroanatomy, physiology, and function of the mammalian brain, as well as the controversies and competing theories in the respective disciplines are all substantial topics on their own right.

Our omission of these issues is not out of lack of recognition or belief that they are unimportant. Rather, our aim in this article has been to present a high-level overview of a richly interdisciplinary set of questions whose broad outlines have only recently begun to take shape. We will tackle these issues and others in a subsequent series of manuscripts and invite interested researchers to join us. Our fundamental motivation in proposing this framework is to bring together scholars from diverse communities that may not be aware of each other's research and their potential for synergy. We believe that there is a wealth of existing research which can be fruitfully re-examined and re-conceptualized from the perspective of artificial intelligence and the value alignment problem. We hope that additional interaction between these communities will help to

[8] In a recent article, Baum has argued that the normative basis for "social choice" and "bottom-up" approaches to AI ethics must overcome strong obstacles that have been insufficiently explored by the AI safety community [43]. Although the approach we describe here decomposes values into more fundamental components, it is not *a priori* in opposition to top-down ethics. In an extreme case, one could certainly imagine employing a purely predetermined approach to ethics within the context of mammalian values in which no value learning takes place. However, as we stated above, we suspect that an intermediate ground will be found when the issues are more thoroughly examined, and for that reason, we are reluctant to endorse either a bottom-up or a top-down approach too strongly. Given the intellectual youth of the field of AI safety, we see little reason to give strong preference to one set of approaches over the other. Moreover, an important observation that Baum makes in framing his argument is that considerable work relevant to AI ethics already exists in the social choice literature, and yet none of this work has been discussed in any detail by the AI safety community. In our minds, this is a more fundamental point, namely, that there is substantial scholarship in many areas of academic research relevant to AI safety. For this reason, we believe that where there is controversy, the first step should be to ensure that the best possible representations of given viewpoints have been made visible and adequately discussed before endorsing particular courses of action.

refine and more precisely define research problems relevant to designing safe AI goal structures.

Acknowledgements

We would like to thank Adam Safron, Owain Evans, Daniel Dewey, Miles Brundage, and several anonymous reviewers for insightful discussions and feedback on the manuscript. We would also like to thank the guest editors of *Informatica*, Ryan Carey, Matthijs Maas, Nell Watson, and Roman Yampolskiy, for organizing this special issue.

References

[1] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice Hall Press, 3rd ed., 2009.

[2] N. J. Nilsson, *The Quest for Artificial Intelligence*. Cambridge University Press, 2009.

[3] M. Andreessen, "Why Software Is Eating The World," *Wall Street Journal*, vol. 20, 2011.

[4] V. C. Müller and N. Bostrom, "Future Progress in Artificial Intelligence: A survey of expert opinion," in *Fundamental issues of artificial intelligence*, pp. 553–570, Springer, 2016.

[5] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "When Will AI Exceed Human Performance? Evidence from AI Experts," *ArXiv e-prints*, May 2017.

[6] R. Brooks, "The Seven Deadly Sins of AI Predictions," *MIT Technology Review*, vol. 10, no. 6, 2017.

[7] D. Deutsch, "How close are we to creating artificial intelligence?," *AEON Magazine*, vol. 10, no. 3, 2012.

[8] J. Somers, "The Man Who Would Teach Machines to Think," *The Atlantic*, vol. 11, 2013.

[9] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. OUP Oxford, 2014.

[10] M. Shanahan, *The Technological Singularity*. MIT Press, 2015.

[11] I. J. Good, "Speculations Concerning the First Ultraintelligent Machine," *Advances In Computers*, vol. 6, no. 99, pp. 31–83, 1965.

[12] D. Chalmers, "The Singularity: A Philosophical Analysis," *Journal of Consciousness Studies*, vol. 17, no. 9-10, pp. 7–65, 2010.

[13] E. Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in *Global Catastrophic Risks* (Nick Bostrom and Milan Cirkovic, ed.), p. 303, Oxford University Press Oxford, UK, 2008.

[14] S. Russell, "Should We Fear Supersmart Robots?," *Scientific American*, vol. 314, no. 6, pp. 58–59, 2016.

[15] S. Omohundro, "Autonomous technology and the greater human good," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 26, no. 3, pp. 303–315, 2014.

[16] S. M. Omohundro, "The Basic AI Drives," in *AGI*, vol. 171, pp. 483–492, 2008.

[17] A. Y. Ng and S. J. Russell, "Algorithms For Inverse Reinforcement Learning," in *International Conference on Machine Learning*, pp. 663–670, 2000.

[18] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "Cooperative inverse reinforcement learning," 2016.

[19] C. L. Baker, R. R. Saxe, and J. B. Tenenbaum, "Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution," in *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society*, pp. 2469–2474, 2011.

[20] O. Evans, A. Stuhlmüller, and N. D. Goodman, "Learning the Preferences of Ignorant, Inconsistent Agents," *arXiv:1512.05832*, 2015.

[21] O. Evans and N. D. Goodman, "Learning the Preferences of Bounded Agents," in *NIPS Workshop on Bounded Optimality*, 2015.

[22] M. O. Riedl and B. Harrison, "Using Stories to Teach Human Values to Artificial Agents," in *Proceedings of the 2nd International Workshop on AI, Ethics and Society, Phoenix, Arizona*, 2016.

[23] M. O. Riedl, "Computational Narrative Intelligence: A Human-Centered Goal for Artificial Intelligence," *arXiv preprint arXiv:1602.06484*, 2016.

[24] R. Fisher and W. Ury, *Getting to Yes*. Simon & Schuster Sound Ideas, 1987.

[25] I. Horswill, "Men Are Dogs (and Women Too)," in *AAAI Fall Symposium: Naturally-Inspired Artificial Intelligence*, pp. 67–71, 2008.

[26] L. W. Swanson, "Cerebral Hemisphere Regulation of Motivated Behavior," *Brain Research*, vol. 886, no. 1, pp. 113–164, 2000.

[27] L. W. Swanson, *Brain Architecture: Understanding the Basic Plan*. Oxford University Press, 2012.

[28] J. H. Barkow, L. Cosmides, and J. Tooby, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, 1995.

[29] S. Dehaene and L. Cohen, "Cultural Recycling of Cortical Maps," *Neuron*, vol. 56, no. 2, pp. 384–398, 2007.

[30] C. Peterson and M. E. Seligman, *Character Strengths and Virtues: A Handbook and Classification*. Oxford University Press, 2004.

[31] S. Schnall, J. Haidt, G. L. Clore, and A. H. Jordan, "Disgust as Embodied Moral Judgment," *Personality and social psychology bulletin*, 2008.

[32] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, "How to Grow a Mind: Statistics, Structure, and Abstraction," *Science*, vol. 331, no. 6022, pp. 1279–1285, 2011.

[33] J. Bowlby, *Attachment and Loss*, vol. 3. Basic books, 1980.

[34] S. W. Porges, "Orienting in a Defensive World: Mammalian Modifications of Our Evolutionary Heritage. A Polyvagal Theory," *Psychophysiology*, vol. 32, no. 4, pp. 301–318, 1995.

[35] J. Cassidy, *Handbook of Attachment: Theory, Research, and Clinical Applications*. Rough Guides, 2002.

[36] M. Tomasello, *The Cultural Origins of Human Cognition*. Harvard University Press, 1999.

[37] J. Panksepp and L. Biven, *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*. WW Norton & Company, 2012.

[38] R. List, "Why I Identify as a Mammal," *The New York Times*, 10 2015.

[39] J. Panksepp, *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford university press, 1998.

[40] S. Armstrong and J. Leike, "Towards Interactive Inverse Reinforcement Learning," in *NIPS*, 2016.

[41] J. Greene and J. Haidt, "How (and where) does moral judgment work?," *Trends in Cognitive Sciences*, vol. 6, no. 12, pp. 517–523, 2002.

[42] J. D. Greene, "The cognitive neuroscience of moral judgment," *The Cognitive Neurosciences*, vol. 4, pp. 1–48, 2009.

[43] S. D. Baum, "Social Choice Ethics in Artificial Intelligence," *AI & SOCIETY*, pp. 1–12, 2017.

Robust Computer Algebra, Theorem Proving, and Oracle AI

Gopal P. Sarma
School of Medicine, Emory University, Atlanta, GA USA
E-mail: gopal.sarma@emory.edu

Nick J. Hay
Vicarious FPC, San Francisco, CA USA
E-mail: nnickhay@gmail.com

Keywords: Oracle AI, AI safety, CAS, theorem proving, math oracles

Received: June 24, 2013

In the context of superintelligent AI systems, the term "oracle" has two meanings. One refers to modular systems queried for domain-specific tasks. Another usage, referring to a class of systems which may be useful for addressing the value alignment and AI control problems, is a superintelligent AI system that only answers questions. The aim of this manuscript is to survey contemporary research problems related to oracles which align with long-term research goals of AI safety. We examine existing question answering systems and argue that their high degree of architectural heterogeneity makes them poor candidates for rigorous analysis as oracles. On the other hand, we identify computer algebra systems (CASs) as being primitive examples of domain-specific oracles for mathematics and argue that efforts to integrate computer algebra systems with theorem provers, systems which have largely been developed independent of one another, provide a concrete set of problems related to the notion of provable safety that has emerged in the AI safety community. We review approaches to interfacing CASs with theorem provers, describe well-defined architectural deficiencies that have been identified with CASs, and suggest possible lines of research and practical software projects for scientists interested in AI safety.

Povzetek: Obravnavani so raziskovalni problemi, povezani z računskimi sistemi s prerokom in varnostjo umetne inteligence.

1 Introduction

Recently, significant public attention has been drawn to the consequences of achieving human-level artificial intelligence. While there have been small communities analyzing the long-term impact of AI and related technologies for decades, these forecasts were made before the many recent breakthroughs that have dramatically accelerated the pace of research in areas as diverse as robotics, computer vision, and autonomous vehicles, to name just a few [1–3].

Most researchers and industrialists view advances in artificial intelligence as having the potential to be overwhelmingly beneficial to humanity. Medicine, transportation, and fundamental scientific research are just some of the areas that are actively being transformed by advances in artificial intelligence. On the other hand, issues of privacy and surveillance, access and inequality, or economics and policy are also of utmost importance and are distinct from the specific technical challenges posed by most cutting-edge research problems [4, 5].

In the context of AI forecasting, one set of issues stands apart, namely, the consequences of artificial intelligence whose capacities vastly exceed that of human beings. Some researchers have argued that such a "superintelligence" poses distinct problems from the more modest AI systems described above. In particular, the emerging discipline of AI safety has focused on issues related to the potential consequences of mis-specifying goal structures for AI systems which have significant capacity to exert influence on the world. From this vantage point, the fundamental concern is that deviations from "human-compatible values" in a superintelligent agent could have significantly detrimental consequences [1].

One strategy that has been advocated for addressing safety concerns related to superintelligence is Oracle AI, that is, an AI system that only answers questions. In other words, an Oracle AI does not directly influence the world in any capacity except via the user of the system. Because an Oracle AI cannot directly take physical action except by answering questions posed by the system's operator, some have argued that it may provide a way to bypass the immediate need for solving the "value alignment problem" and would itself be a powerful resource in enabling the safe design of autonomous, deliberative superintelligent agents [1, 6–9].

A weaker notion of the term oracle, what we call a *domain-specific oracle*, refers to a modular component of a larger AI system that is queried for domain-specific tasks. In this article, we view computer algebra systems as primitive domain-specific oracles for mathematical computation

which are likely to become quite powerful on the time horizons on which many expect superintelligent AI systems to be developed [10, 11]. Under the assumption that math oracles prove to be useful in the long-term development of AI systems, addressing well-defined architectural problems with CASs and their integration with interactive theorem provers provides a concrete set of research problems that align with long-term issues in AI safety. In addition, such systems may also be useful in proving the functional correctness of other aspects of an AI architecture. In Section 2, we briefly discuss the unique challenges in allocating resources for AI safety research. In Section 3, we briefly summarize the motivation for developing oracles in the context of AI safety and give an overview of safety risks and control strategies which have been identified for superintelligent oracle AIs. In Section 4 we analyze contemporary question answering systems and argue that in contrast to computer algebra systems, current consumer-oriented, NLP-based systems are poor candidates for rigorous analysis as oracles. In Section 5, we review the differences between theorem provers and computer algebra systems, efforts at integrating the two, and known architectural problems with CASs. We close with a list of additional research projects related to mathematical computation which may be of interest to scientists conducting research in AI safety.

2 Metascience of AI safety research

From a resource allocation standpoint, AI safety poses a unique set of challenges. Few areas of academic research operate on such long and potentially uncertain time horizons. This is not to say that academia does not engage in long-term research. Research in quantum gravity, for example, is approaching nearly a century's worth of effort in theoretical physics [12]. However, the key difference between open-ended, fundamental research in the sciences or humanities and AI safety is the possibility of negative consequences, indeed significant ones, of key technological breakthroughs taking place without corresponding advances in frameworks for safety [1, 13] .

These issues have been controversial, largely due to disagreement over the time-horizons for achieving human-level AI and the subsequent consequences [10, 11]. Specifically, the notion of an "intelligence explosion," whereby the intelligence of software systems dramatically increases due their capacity to model and re-write their own source code, has yet to receive adequate scientific scrutiny and analysis [14].

We affirm the importance of AI safety research and also agree with those who have cautioned against proceeding down speculative lines of thinking that lack precision. Our perspective in this article is that it is possible to fruitfully discuss long-term issues related to AI safety while maintaining a connection to practical research problems. To some extent, our goal is similar in spirit to the widely discussed manuscript "Concrete Problems in AI Safety" [15]. However, we aim to be a bit more bold. While the authors of "Concrete Problems" state at the outset that their analysis will set aside questions related to superintelligence, our goal is to explicitly tackle superintelligence related safety concerns. We believe that there are areas of contemporary research that overlap with novel ideas and concepts that have arisen among researchers who have purely focused on analyzing the consequences of AI systems whose capacities vastly exceed those of human beings.

To be clear, we do not claim that the strategy of searching for pre-existing research objectives that align with the aims of superintelligence theory is sufficient to cover the full spectrum of issues identified by AI safety researchers. There is no doubt that the prospect of superintelligence raises entirely new issues that have no context in contemporary research. However, considering how young the field is, we believe that the perspective adopted in this article is a down-to-earth and moderate stance to take while the field is in a critical growth phase and a new culture is being created.

This article focuses on one area of the AI safety landscape, Oracle AI. We identify a set of concrete software projects that relate to more abstract, conceptual ideas from AI safety, to bridge the gap between practical contemporary challenges and longer term concerns which are of an uncertain time horizon. In addition to providing concrete problems for researchers and engineers to tackle, we hope this discussion will be a useful introduction to the concept of Oracle AI for newcomers to the subject. We state at the outset that within the context of Oracle AI, our analysis is limited in scope to systems which perform mathematical computation, and not to oracles in general. Nonetheless, considering how little effort has been directed at the superintelligence control problem, we are confident that there is low-hanging fruit in addressing these more general issues which are awaiting discovery.

3 Brief overview of Oracle AI

As described above, an Oracle AI is a system which only answers questions. Although the term has traditionally been used to denote fully general AI systems, we can also imagine oracles whose capabilities are more limited. The latter systems are what we refer to as *domain-specific oracles*. AI safety researchers have argued that superintelligent oracles have the potential to help address safety concerns of autonomous, deliberative superintelligent AI systems, as an Oracle AI's only ability to influence the world is through the user of the system. Consequently, an Oracle AI itself may be in a position to assist researchers in solving difficult technical and conceptual problems related to the design of independent, deliberative agents. Yet superintelligent oracle AIs pose their own distinct set of safety concerns. The problems identified in [6], for example, presume a superintelligent AI system, which while limited in its ability to

interact with the world, may nonetheless have the potential to negatively impact the world by manipulating the user of the system. As Armstrong, Bostrom, and Sandberg write:

> We are all vulnerable to the OAI's social engineering. An advanced OAI can literally offer us the world. Genuine altruism is no defence against this; it can offer a world free of poverty, hunger, war, disease or death, as well as all personal trinkets we could want. It could guess the weaknesses of each individual, and find the right arguments to convince us that granting it power or liberty is the moral and profitable thing to do. It will be able to play on rivalries, institutional weaknesses and ethical dilemmas with more virtuosity than any human could. [6]

Given the potential dangers of human manipulation by an OAI, researchers have searched for control strategies to mitigate such risks. A few examples of control methods that have been proposed are:

- **Physical security**- physically containing the OAI system
- **Limiting or eliminating external manipulators**- like the previous security measure, precluding the OAI from physically manipulating the world in any direct manner
- **Epistemic capability control**- the OAIs output could be restricted even further beyond question answering, for example, restricting it to answer "yes / no " questions
- **Virtual worlds**- the OAI could be run in a simulated world so that even if the system's capabilities were expanded, its behavior could be observed and modeled.
- **Resetting the OAI after each question**- this would prevent the OAI from engaging in long-term social engineering by manipulating the answers to the users' questions

Although the capacities of domain-specific oracles are more limited, they can nonetheless pose safety risks. Architectural deficiencies in such oracles might be exploited by a larger AI system to manipulate the human user. It could give answers which are difficult to verify and which allow the OAI to execute complex and intricate plans unbeknownst to the user. Therefore, while flaws in domain-specific oracles are not inherently risky if used solely in their domain of applicability, they may very well be dangerous as part of a larger system with more general capabilities. Though not a "control strategy" in the narrowest sense, creating "robust" domain-specific oracles is an important objective in designing safe OAIs. Furthermore, ensuring the robustness of domain-specific subsystems might mitigate the need for stronger control strategies, as the OAI would have fewer weaknesses to exploit.

It should go without saying that the arguments presented above are highly schematic and do not dependent on specific technologies. To our knowledge, there is very limited work on translating analyses of superintelligent oracle AIs into the concrete language of modern artificial intelligence [8, 9, 16]. Our goal in this manuscript is in this spirit, that is, to anchor schematic, philosophical arguments in practical, contemporary research. To do so, we will narrow our focus to the mathematical domain. In the remainder of the article, we will use the term oracle in the more limited sense of a domain-specific subsystem, and in particular, oracles for performing mathematical computations. We hope that the analysis presented here will be of intrinsic value in developing robust math oracles, as well as provide some intuition and context for identifying concrete problems relevant to developing safe, superintelligent oracle AI systems.

4 Are there contemporary systems which qualify as oracles?

The obvious class of contemporary systems which would seem to qualify as oracles are question answering systems (QASs). As we stated above, a basic criterion characterizing oracles is that their fundamental mode of interaction is answering questions posed by a user, or for domain-specific queries as part of a larger AI system.

Contemporary QASs are largely aimed at using natural language processing techniques to answer questions pertaining to useful facts about the world such as places, movies, historical figures, and so on. An important point to make about QASs is the highly variable nature of the underlying technology. For instance, IBM's original Watson system which competed in Jeopardy, was developed prior to the recent advances in deep learning which have fundamentally transformed areas ranging from computer vision, to speech recognition, to natural language processing [17]. In this particular task, the system was nonetheless able to perform at a level beyond that of the most accomplished human participants. The introduction of "info panes" into popular search engines, on the other hand, have been based on more recent machine learning technology, and indeed, these advances are also what power the latest iterations of the Watson system [18]. On the other end of the spectrum is Wolfram | Alpha, which is also a question answering system, but which is architecturally centered around a large, curated repository of structured data, rather than datasets of unstructured natural language [19].

While these systems are currently useful for humans in navigating the world, planning social outings, and arriving at quick and useful answers to ordinary questions, it is not clear that they will remain useful in quite the same capacity many years from now, or as standalone components of superintelligent AI systems. Although the underlying techniques of deep learning or NLP are of fundamental interest in their own right, the fact that these systems are QASs at

all seems to be more of an artifact of their utility for consumers.

Another important observation about contemporary QASs is that much of their underlying NLP-based architecture can be replaced by taking advantage of structured data, as the example of Wolfram — Alpha demonstrates. For the other NLP or machine learning based systems, the underlying technology can be used as part of larger, semi-automated pipelines to turn unstructured data from textual sources into structured data. Once again, this fact simply underscores that contemporary QASs are not particularly appealing model systems to analyze from the Oracle AI safety perspective.[1]

4.1 Computer algebra and domain-specific oracles for mathematical computation

The question answering systems described above all rely on natural language processing to varying degrees. In addition, their domain of applicability has tended towards "ordinary" day-to-day knowledge useful to a wide array of consumers. Another type of question answering system is a computer algebra system (CAS). Computer algebra has traditionally referred to systems for computing specific results to specific mathematical equations, for example, computing derivatives and integrals, group theoretic quantities, etc. In a sense, we can think of computer algebra as a set of algorithms for performing what an applied mathematician or theoretical physicist might work out on paper and pencil. Indeed, some of the early work in computer algebra came from quantum field theory—one of the first computer algebra systems was Veltman's *Schoonschip* for performing field theoretic computations that led to the theory of electroweak unification [20].

As computer algebra systems have grown in popularity, their functionality has expanded substantially to cover a wide range of standard computations in mathematics and theoretical physics, including differentiation, integration, matrix operations, manipulation of symbolic expressions, symbolic substitution, algebraic equation solving, limit computation, and many others. Computer algebra systems typically run in a `read, evaluate, print` loop (`repl`), and in the research and education context, their popularity has also grown as a result of the notebook model pioneered by the *Mathematica* system, allowing for computations in CASs to closely mimic the sequential, paper and pencil work of mathematicians and theoretical physicists.

In assessing the long-term utility of CASs, it is important to note that there is little reason to believe that computer algebra will be subsumed by other branches of AI research such as machine learning. Indeed, recent research has demonstrated applications of machine learning to both computer algebra and theorem proving (which we discuss in more detail below), via algorithm selection in the former case [21] and proof assistance in the latter [22, 23]. While certainly not as visible as machine learning, computer algebra and theorem proving are very much active and deep areas of research which are also likely to profit from advances in other fields of artificial intelligence, as opposed to being replaced by them [24]. On the time horizons on which we are likely to see human-level artificial intelligence and beyond, we can expect that these systems will become quite powerful, and possess capabilities that may be useful in the construction of more general AI systems. Therefore, it is worth examining such systems from the perspective of AI safety.

4.2 Briefly clarifying nomenclature

Before proceeding, we want to explicitly describe issues relating to nomenclature that have arisen in the discussion thus far, and state our choices for terminology. Given that the phrase "Oracle AI" has become common usage in the AI safety community, we will continue to use this phrase, with the first word capitalized, as well as the acronym OAI. Where clarification is needed, we may also use the full phrase "superintelligent oracle AI," without capitalization.

For more modest use cases of the word oracle, we will either refer to "domain-specific oracles," or state the domain of knowledge where the oracle is applicable. We can, at the very least in the abstract, consider extending this terminology to other domains such as "physics oracles," "cell biology oracles," or "ethics oracles" and so on. Therefore, the remainder of the article will be concerned with safety and robustness issues in the design of "math oracles."

5 Robust computer algebra and integrated theorem proving

Today we should consider as a standard feature much closer interaction between proof assistance and computer algebra software. Several areas can benefit from this, including specification of interfaces among components, certification of results and domains of applicability, justification of optimizations and, in the other direction, use of efficient algebra in proofs.

[1] We emphasize that our argument that contemporary QASs are not good candidates for analysis as Oracle AIs is not an argument against the traditional formulation of Oracle AI as a tool for AI safety. We fully expect significant breakthroughs to be made in advancing the theory and practice of oracle-based techniques for AI safety and we hope that this manuscript will provide some motivation to pursue such research. Rather, our point is that when viewing contemporary systems from the lens of superintelligence, there seems little reason to believe that current NLP-based QASs will remain sufficiently architecturally stable to be used as standalone components in AI systems many years from now. On the other hand, there are certainly important *present-day* problems to examine when evaluating the broader impact of QASs, such as bias in NLP systems, overgeneralization, and privacy, to name just a few. Some of these issues overlap with the set of problems identified in [15] as examples of concrete problems in AI safety. In addition, we are beginning to see conferences devoted to contemporary ethical issues raised by machine learning. See, for example, the workshop Ethics in Natural Language Processing (https://www.aclweb.org/portal/content/first-workshop-ethics-natural-language-processing).

> - **Stephen Watt in *On the future of computer algebra systems at the threshold of 2010***

As we described above, computer algebra systems can be thought of as question answering systems for a subset of mathematics. A related set of systems are interactive proof assistants or interactive theorem provers (ITPs). While ITPs are also systems for computer-assisted mathematics, it is for a different mathematical context, for computations in which one wishes to construct a proof of a general kind of statement. In other words, rather than computing specific answers to specific questions, ITPs are used to show that candidate mathematical structures (or software systems) possess certain properties.

In a sense, the distinction between theorem proving and computer algebra should be viewed as a historical anomaly. From the perspective of philosophical and logical efforts in the early 20th century that led to the "mechanization of mathematics" the distinction between computing the n^{th} Laguerre polynomial and constructing a proof by induction might have been viewed as rather artificial, although with the benefit of hindsight we can see that the two types of tasks are quite different in practice [25].

The role of ITPs in the research world is very different from that of CASs. Whereas CASs allow researchers to perform difficult computations that would be impossible with paper and pencil, constructing proofs using ITPs is often more difficult than even the most rigorous methods of pure mathematics. In broad terms, the overhead of using ITPs to formalize theorems arises from the fact that proofs in these systems must proceed strictly from a set of formalized axioms so that the system can verify each computation. Consequently, ITPs (and related systems, such as automatic theorem provers) are largely used for verifying properties of mission-critical software systems which require a high-degree of assurance, or for hardware verification, where mistakes can lead to costly recalls [26–30].

As the quotation above suggests, many academic researchers view the integration of interactive proof assistants and computer algebra systems as desirable, and there have been numerous efforts over the years at exploring possible avenues for achieving this objective [31–34] (a more complete list is given below). By integrating theorem proving with computer algebra, we would be opening up a wealth of potentially interoperable algorithms that have to date remained largely unintegrated. To cite one such example, in [35], the authors have developed a framework for exchange of information between the Maple computer algebra system and the Isabelle interactive theorem prover. They show a simple problem involving the proof of an elementary polynomial identity that could be solved with the combined system, but in neither system alone (see Fig. 1).

We cite this example to demonstrate how a simply stated elementary problem cannot be solved in existing environments for either computer algebra or proof assistance. The computer algebra system does not have the capacity for structural induction and theorem provers generally have rather weak expression simplifiers. There are numerous examples such as this one in the academic literature.

Another key difference between CASs and ITPs is the architectural soundness of the respective systems. As we will discuss below, computer algebra systems have well-defined architectural deficiencies, which while not a practical issue for the vast majority of use cases, pose problems for their integration with theorem provers, which by their nature, are designed to be architecturally sound. In the context of superintelligent AI systems, the architectural problems of CASs are potential points of weakness that could be exploited for malicious purposes or simply lead to unintended and detrimental consequences. Therefore, we use the phrase "robust computer algebra" to refer to CASs which lack the problems that have been identified in the research literature. In the section below, we combine the discussion of robust computer algebra and integration with interactive theorem provers, as there is a spectrum of approaches which address both of these issues to varying degrees.

5.1 A taxonomy of approaches

There are many possible avenues to tackle the integration of theorem provers with computer algebra systems. We give 4 broad categories characterizing such integration efforts[2]:

1. **Theorem provers built on top of computer algebra systems:** These include Analytica, Theorema, RedLog, and logical extensions to the Axiom system [34, 36–39] .

2. **Frameworks for mathematical exchange between the two systems:** This category includes MathML, OpenMath, OMSCS, MathScheme, and Logic Broker [40–44].

3. **"Bridges" or "ad-hoc" information exchange solutions:** The pairs of systems in this category include bridges combining PVS, HOL, or Isabelle with Maple, NuPRL with Weyl, Omega with Maple/GAP, Isabelle with Summit, and most recently, Lean with *Mathematica* [35, 45–51]. The example given above, bridging Isabelle and Maple, is an example of an approach from this category.

4. **Embedding a computer algebra system inside a proof assistant:** This is the approach taken by Kaliszyk and Wiedijk in the HOLCAS system. In their system, all expressions have precise semantics, and the proof assistant proves the correctness of each simplification made by the computer algebra system [32].

One primary aspect of integration that differentiates these approaches is the degree of trust the theorem prover places in the computer algebra system. Computer algebra

[2]This classification was first described by Kaliszyk and Wiedijk [32] in a paper arguing for an architecture which we list as the fourth category given above.

$$
\begin{array}{l}
1 : TH \vdash_I n^5 \leq 5^n \\
2 : TH \vdash_I 5^5 \leq 5^5 \\
3 : TH \vdash_I n \leq 5 \\
4 : TH \vdash_I \forall x : [x \in N \wedge 5 \leq x \wedge x^5 \leq 5^x] \Longrightarrow (x+1)^5 \leq 5^{(x+1)} \\
5 : x \in N \vdash_M (x+1)^5 \equiv x^5 + 5x^4 + 10x^3 + 10x^2 + 5x + 1 \\
6 : TH \vdash_I \forall x : [x \in N \wedge 5 \leq x \wedge x^5 \leq 5^x] \Longrightarrow \\
\qquad x^5 + 5x^4 + 10x^3 + 10x^2 + 5x + 1 \leq 5^{(x+1)} \\
7 : x \in N \vdash_M 5^{(x+1)} \equiv 5 * 5^x \\
8 : TH \vdash_I \forall x : [x \in N \wedge 5 \leq x \wedge x^5 \leq 5^x] \Longrightarrow \\
\qquad x^5 + 5x^4 + 10x^3 + 10x^2 + 5x + 1 \leq 5 * 5^x
\end{array}
$$

Figure 1: Example of a polynomial identity proven by integrating the Maple computer algebra system with Isabelle. Maple's simplifier is used for expanding polynomials—a powerful complement to the theorem proving architecture of Isabelle which allows for the setup of a proof by induction.

systems give the false impression of being monolithic systems with globally well-defined semantics. In reality, they are large collections of algorithms which are neatly packaged into a unified interface. Consequently, there are often corner cases where the lack of precise semantics can lead to erroneous solutions. Consider the following example:

```
(%i1) equations:[(x-1)^2/(x^2-1)];
                                        2
                                   (x - 1)
(%o1)                             [--------]
                                     2
                                    x  - 1
(%i2) solutionsn:solve(equations, [x]);
(%o2)                              [x = 1]
```

Figure 2: Example of an incorrect solution to a simple polynomial equation by a computer algebra system.

The system incorrectly gives 1 as a solution, even though the given polynomial has an indeterminate value for $x = 1$. However, because the expression is treated as a fraction of polynomials, it is first simplified before the solve operation is applied. In other words, there is an unclear semantics between the solver module and the simplifier which leads to an incorrect result.

Another simple example is the following integral:

$$\text{In[3]:=} \quad \int x^n \, dx$$

$$\text{Out[3]=} \quad \frac{x^{1+n}}{1+n}$$

Figure 3: A problem arising in symbolic integration due to the non-commutativity of evaluation and substitution.

Making the substitution $n = -1$ gives an indeterminate result, while it is clear by inspection that the solution to the integral for $n = -1$ is simply $ln(x)$. This belongs to a class of problems known as the *specialization problem*, namely that expression evaluation and variable substitution do not commute [31]. So while we have seen above that theorem proving can benefit tremendously from the wealth of algorithms for expression simplification and mathematical knowledge in computer algebra, there is the potential cost of compromising the reliability of the combined system. As a possible application to current research in AI safety, consider the decision-theoretic research agenda for the development of safe, superintelligent AI systems outlined in [52–56]. If we require formal guarantees of correctness at any point in a sequence of computations in which computer algebra is used, current systems would be unable to provide the necessary framework for constructing such a proof.

5.1.1 Qualitatively certified computations

In our taxonomy of approaches to bridging theorem provers with computer algebra, we described how a key distinction was the degree of trust that the theorem prover places in the computer algebra system. For instance, approaches which build theorem provers on top of computer algebra systems do not address the architectural issues with CASs. They are integrative, but not more sound. On the other extreme, building a computer algebra system on top of a theorem prover allows for a degree of trust that is on par with that of the theorem prover itself. However, this approach has the distinct disadvantage that computer algebra systems represent many hundred man-years worth of effort.

The more intermediate approaches involving common languages for symbolic exchange or ad-hoc bridges, bring to light an important notion in the spectrum of provable sa-

fety, namely the ability to assign probabilities for the correctness of computations. In [57], the authors present an algorithm for assigning probabilities to any statement in a formal language. We might ask what strategies might look like that have a similar goal in mind, but are significantly weaker. Interfaces between theorem provers and computer algebra systems provide a concrete example where we can ask a question along these lines. Fundamentally, in such an interface, the computer algebra system is the weaker link and should decrease our confidence in the final result. But by how much? For instance, in the example given in Figure 1, how should we revise our confidence in the result knowing that polynomial simplification was conducted within a computer algebra system?

It is worth asking for simple answers to this question that do not require major theoretical advances to be made. For instance, we might imagine curating information from computer algebra experts about known weaknesses, and use this information to simply give a qualitative degree of confidence in a given result. Or, for example, in a repository of formal proofs generated using integrated systems, steps of the proof that require computer algebra can be flagged and also assigned a qualitative measure of uncertainty.

The relationship that this highly informal method of giving qualitative certification to computations has with the formal algorithm developed in [57] can be compared to existing techniques in the software industry for ensuring correctness. On the one hand, unit testing is a theoretically trivial, yet quite powerful practice, something along the lines of automated checklists for software. The complexities of modern software would be impossible to handle without extensive software testing frameworks [58–62]. On the other hand, formal verification can provide substantially stronger guarantees, yet is a major undertaking, and the correctness proofs are often significantly more demanding to construct than the software itself. Consequently, as discussed in Section 5, formal verification is much less frequently used in industry, typically only in exceptional circumstances where high guarantees of correctness are required, or for hardware verification [26–30].

Integrated systems for computer algebra and theorem proving give rise to a quite interesting (and perhaps ironic) opportunity to pursue simple strategies for giving qualitative estimates for the correctness of a computation.

5.1.2 Logical failures and error propagation

As the examples described above demonstrate, errors in initial calculations may very well propagate and give rise to non-sensical results. As AI systems capable of performing mathematical computation become increasingly sophisticated and embedded as part of design workflows for science and engineering (beyond what we see today), we could imagine such errors being quite costly and difficult to debug. In the case of a superintelligent AI system, more concerning scenarios would be if systematic errors in computer algebra could be exploited for adversarial purposes or if they led to unintentional accidents on a large scale.

The issue of error propagation is another example of a concrete context for pursuing simple strategies for assigning qualitative measures of certainty to computations performed by integrated theorem proving / computer algebra systems. For instance, we may be less inclined to trust a result in which the computer algebra system was invoked early on in a computation as opposed to later. With curated data from computer algebra experts on the reliability or failure modes of various algorithms, we might also chain together these informal estimates to arrive at a single global qualitative estimate. If multiple systems were to be developed independently, or which were based on fundamentally different architectures, we might also be significantly more confident in a result which could be verified by two separate systems.

5.1.3 Additional topics

Some related ideas merit investigation in the broader context of mathematical computation:

- **Integrating SMT solvers with interactive theorem provers:** Satisfiability modulo theories (SMT) solvers are an important element of automated reasoning and there have been efforts analogous to those described above to bridge SMT solvers with interactive theorem provers [63, 64].
- **Identifying the most important / widely used algorithms in computer algebra:** Computer algebra systems have grown to become massive collections of algorithms extending into domains well outside of the realm of mathematics. If the purely mathematical capacities of CASs prove to be useful in future AI systems, it would be valuable to rank order algorithms by their popularity or importance.

  One approach would be to do basic textual analysis of the source code from GitHub or StackExchange. This would also allow for more targeted efforts to directly address the issues with soundness in core algorithms such as expression simplification or integration. In the context of the HOLCAS system described above, for example, it would be valuable to have rough estimates for the number of man-hours required to implement a minimal CAS with the most widely used functionality on top of a theorem prover.
- **Proof checkers for integrated systems:** Proof checkers are important tools in the landscape of formal verification and theorem proving. Indeed, as it is often much less computationally expensive to verify the correctness of a proof than to generate it from scratch, the availability of proof checkers for the widely used interactive theorem provers is one reason we can be confident in the correctness of formal proofs [65, 66].

  As we described above, strategies for integrating computer algebra with theorem provers can potentially result in a combined system which is less trustworthy

than the theorem prover alone. Therefore, the availability of proof checkers for combined systems would be a valuable resource in verifying proof correctness, and in certain mathematical domains, potentially provide an avenue for surmounting the need to directly make the CAS itself more architecturally robust.

The development of integrated proof checkers is likely to be a substantial undertaking and require novel architectures for integrating the core CAS and ITP systems distinct from what has been described above. However, it is a largely unexplored topic that merits further investigation.

- **Analyzing scaling properties of algorithms for computer algebra and theorem proving as a function of hardware resources:** The premise of the analysis presented above is that CASs (and integrated theorem proving) are likely to remain sufficiently architecturally stable and useful on a several decade time-horizon in the construction of AI systems. On the other hand, as we argued earlier, it is much less clear that the same will be true of the most visible, NLP-based, consumer-oriented question answering systems. To make these arguments more rigorous, it would be valuable to develop quantitative predictions of what the capabilities will be of existing algorithms for computer algebra and theorem proving when provided with substantially expanded hardware resources. For instance, we might examine problems in mathematics or theoretical physics for which naïve solutions in CASs are intractable with current resources, but which may be feasible with future hardware.

- **The cognitive science of computer algebra:** What role has computer algebra played in theoretical physics and mathematics? How has it influenced the thinking process of researchers? Has computer algebra simply been a convenience that has shifted the way problems are solved, or has it fundamentally enabled new problems to be solved that would have been completely intractable otherwise?

  The cognitive science of mathematical thought is a substantial topic which overlaps with many established areas of research [67–71]. However, a systematic review of research in mathematics and theoretical physics since the advent of computer algebra and its role in the mathematical thought process is an underexplored topic. It would be an interesting avenue to pursue in understanding the role that CASs, ITPs, and integrated systems may come to play in superintelligence, particularly in the case of neuromorphic systems that have been modeled after human cognition. These questions also relate to understanding the scaling properties of CAS and theorem proving algorithms as well as cataloguing the most widely used algorithms in computer algebra.

6 Conclusion

The aim of this article has been to examine pre-existing research objectives in computer science and related disciplines which align with problems relevant to AI safety, thereby providing concrete, practical context for problems which are otherwise of a longer time horizon than most research. In particular, we focused on the notion of "Oracle AI" as used in the AI safety community, and observed that the word oracle has two meanings in the context of superintelligent AI systems. One usage refers to a subsystem of a larger AI system queried for domain-specific tasks, and the other to superintelligent AI systems restricted to only answer questions.

We examined contemporary question answering systems (QASs) and argued that due to their architectural heterogeneity, consumer-oriented, NLP-based systems do not readily lend themselves to rigorous analysis from an AI safety perspective. On the other hand, we identified computer algebra systems (CASs) as concrete, if primitive, examples of domain-specific oracles. We examined well-known architectural deficiencies with CASs identified by the theorem proving community and argued that the integration of interactive theorem provers (ITPs) with CASs, an objective that has been an area of research in the respective communities for several decades, provides a set of research problems and practical software projects related to the development of powerful and robust math oracles on a multi-decade time horizon. Independent of their role as domain-specific oracles, such systems may also prove to be useful tools for AI safety researchers in proving the functional correctness of other components of an AI architecture. Natural choices of systems to use would be interfaces for the Wolfram Language, the most widely used computer algebra system, with one of the HOL family of theorem provers or Coq, both of which have substantial repositories of formalized proofs [72–75], or a more modern ITP such as Lean [51, 76].

Rather than representing a bold and profound new agenda, we view these projects as being concrete and achievable goals that may pave the way to more substantial research directions. Because the topics we have discussed have a long and rich academic history, there are a number of "shovel-ready" projects appropriate for students anywhere from undergraduates to PhD students and beyond. Good undergraduate research projects would probably start with some basic data science to catalogue core computer algebra algorithms by their usage and popularity. From there, it would be useful to have an estimate of what certified implementations of these algorithms would entail, whether formally verified implementations, or along the lines of Kaliszyk and Wiedijk's HOLCAS system where the CAS is built on top of a theorem prover. Also useful would be a systematic study of role that computer algebra has played in mathematics and theoretical physics. This would have some interesting overlap with cognitive psychology, and these three projects together would make for an approa-

chable undergraduate thesis, or a beginning project for a graduate student. A solid PhD thesis devoted to the topic of Oracle AI might involve tackling approaches to oracles stemming from reinforcement learning (RL) [8,16], as well as more advanced theorem proving and CAS related topics such as investigating the development of a hybrid architecture that would allow for proof-checking. A student who worked on these projects for several years would develop a unique skill set spanning philosophy, machine learning, theorem proving, and computer algebra.

In the context of superintelligent oracle AIs which may possess the ability to manipulate a human user, we differentiate between addressing architectural or algorithmic deficiencies in subsystems versus general control methods or containment strategies. Given that strong mathematical capabilities are likely to be useful in the construction of more general AI systems, designing robust CASs (and any other domain-specific oracle) is an important counterpart to general control strategies, as the top-level AI system will have fewer loopholes to exploit. Controlling OAIs poses a distinct set of challenges for which concrete mathematical analysis is in its infancy [8,9,16]. Nonetheless, considering how little attention has been given to the superintelligence control problem in general, we are optimistic about the potential to translate the high-level analyses of OAIs that have arisen in the AI safety community into the mathematical and software frameworks of modern artificial intelligence.

Acknowledgements

We would like to thank Stuart Armstrong, David Kristoffersson, Marcello Herreshoff, Miles Brundage, Eric Drexler, Cristian Calude, and several anonymous reviewers for insightful discussions and feedback on the manuscript. We would also like to thank the guest editors of *Informatica*, Ryan Carey, Matthijs Maas, Nell Watson, and Roman Yampolskiy, for organizing this special issue.

References

[1] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[2] M. Shanahan, *The Technological Singularity*. MIT Press, 2015.

[3] D. Chalmers, "The Singularity: A Philosophical Analysis," *Journal of Consciousness Studies*, vol. 17, no. 9-10, pp. 7–65, 2010.

[4] M. Tegmark *et al.*, "An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence (Future of Life Institute)," 2015.

[5] S. Russell, D. Dewey, and M. Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Magazine*, vol. 36, no. 4, pp. 105–114, 2015.

[6] S. Armstrong, A. Sandberg, and N. Bostrom, "Thinking inside the box: Controlling and Using an Oracle AI," *Minds and Machines*, vol. 22, no. 4, pp. 299–324, 2012.

[7] B. Fallenstein, J. Taylor, and P. F. Christiano, "Reflective oracles: A foundation for game theory in artificial intelligence," in *Logic, Rationality, and Interaction*, pp. 411–415, Springer, 2015.

[8] S. Armstrong, "Value and policy networks as Oracle AIs." *in preparation*, 2017.

[9] S. Armstrong, "Good and safe uses of AI Oracles," *ArXiv e-prints*, Nov. 2017.

[10] V. C. Müller and N. Bostrom, "Future Progress in Artificial Intelligence: A survey of expert opinion," in *Fundamental Issues of Artificial Intelligence*, pp. 553–570, Springer, 2016.

[11] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "When Will AI Exceed Human Performance? Evidence from AI Experts," *ArXiv e-prints*, May 2017.

[12] C. Rovelli, "Quantum gravity," *Scholarpedia*, vol. 3, no. 5, p. 7117, 2008.

[13] S. Russell, "Should We Fear Supersmart Robots?," *Scientific American*, vol. 314, no. 6, pp. 58–59, 2016.

[14] A. H. Eden, J. H. Moor, J. H. Soraker, and E. Steinhart, *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Springer Verlag, 2012.

[15] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," *ArXiv e-prints*, June 2016.

[16] S. M. Armstrong and L. Orseau, "Safely Interruptible Agents." *submitted*, 2016.

[17] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, *et al.*, "Building Watson: An overview of the DeepQA project," *AI magazine*, vol. 31, no. 3, pp. 59–79, 2010.

[18] W. Knight, "IBM Pushes Deep Learning with a Watson Upgrade," *MIT Technology Review*, 7 2015.

[19] S. Wolfram, "Jeopardy, IBM, and Wolfram — Alpha," *Stephen Wolfram — Blog*, 1 2011.

[20] S. Weinzierl, "Computer Algebra in Particle Physics," *ArXiv High Energy Physics - Phenomenology e-prints*, Sept. 2002.

[21] Z. Huang, "Machine Learning and Computer Algebra," tech. rep., University of Cambridge, Computer Laboratory, 2016.

[22] G. Irving, C. Szegedy, A. A. Alemi, F. Chollet, and J. Urban, "DeepMath—Deep Sequence Models for Premise Selection," in *Advances in Neural Information Processing Systems*, pp. 2235–2243, 2016.

[23] E. Komendantskaya, J. Heras, and G. Grov, "Machine Learning in Proof General: Interfacing Interfaces," *ArXiv e-prints*, Dec. 2012.

[24] A. Bundy, D. Hutter, C. B. Jones, and J. S. Moore, "AI meets Formal Software Development (Dagstuhl Seminar 12271)," *Dagstuhl Reports*, vol. 2, no. 7, pp. 1–29, 2012.

[25] M. J. Beeson, "The Mechanization of Mathematics," in *Alan Turing: Life and Legacy of a Great Thinker*, pp. 77–134, Springer, 2004.

[26] G. Klein, K. Elphinstone, G. Heiser, J. Andronick, D. Cock, P. Derrin, D. Elkaduwe, K. Engelhardt, R. Kolanski, M. Norrish, *et al.*, "seL4: Formal verification of an OS kernel," in *Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles*, pp. 207–220, ACM, 2009.

[27] R. Kaivola, R. Ghughal, N. Narasimhan, A. Telfer, J. Whittemore, S. Pandav, A. Slobodová, C. Taylor, V. Frolov, E. Reeber, *et al.*, "Replacing Testing with Formal Verification in Intel CoreTM i7 Processor Execution Engine Validation," in *International Conference on Computer Aided Verification*, pp. 414–429, Springer, 2009.

[28] L. Fix, "Fifteen years of formal property verification in Intel," in *25 Years of Model Checking*, pp. 139–144, Springer, 2008.

[29] C. Kern and M. R. Greenstreet, "Formal verification in hardware design: a survey," *ACM Transactions on Design Automation of Electronic Systems*, vol. 4, no. 2, pp. 123–193, 1999.

[30] T. Kropf, *Introduction to Formal Hardware Verification*. Springer Science & Business Media, 2013.

[31] C. Ballarin, *Computer Algebra and Theorem Proving*. PhD thesis, University of Cambridge, Computer Laboratory, 1999.

[32] C. Kaliszyk and F. Wiedijk, "Certified computer algebra on top of an interactive theorem prover," in *Towards Mechanized Mathematical Assistants*, pp. 94–105, Springer, 2007.

[33] S. M. Watt, "On the future of Computer Algebra Systems at the Threshold of 2010," *Proceedings ASCM-MACIS*, pp. 422–430, 2009.

[34] W. Windsteiger, "Theorema 2.0: a system for mathematical theory exploration," in *International Congress on Mathematical Software*, pp. 49–52, Springer, 2014.

[35] P. G. Bertoli, J. Calmet, F. Giunchiglia, and K. Homann, "Specification and integration of theorem provers and computer algebra systems," in *International Conference on Artificial Intelligence and Symbolic Computation*, pp. 94–106, Springer, 1998.

[36] E. Clarke and X. Zhao, "Analytica—A theorem prover in Mathematica," in *International Conference on Automated Deduction*, pp. 761–765, Springer, 1992.

[37] A. Dolzmann and T. Sturm, "Redlog: Computer algebra meets computer logic," *ACM SIGSAM Bulletin*, vol. 31, no. 2, pp. 2–9, 1997.

[38] R. D. Jenks and R. S. Sutor, *AXIOM: The Scientific Computation System*. Springer, 2013.

[39] E. Poll and S. Thompson, "Adding the axioms to Axiom," tech. rep., Computing Laboratory, University of Kent, 1998.

[40] R. Miner, "The importance of MathML to mathematics communication," *Notices of the AMS*, vol. 52, no. 5, pp. 532–538, 2005.

[41] S. Buswell, O. Caprotti, D. P. Carlisle, M. C. Dewar, M. Gaetano, and M. Kohlhase, "The Open Math Standard," tech. rep., The Open Math Society, 2004.

[42] J. Calmet and V. Lefevre, "Toward the Integration of Numerical Computations into the OMSCS Framework," in *7th International Workshop on Computer Algebra in Scientific Computing-CASC*, pp. 71–79, 2004.

[43] J. Carette, W. M. Farmer, and R. O'Connor, "MathScheme: project description," in *International Conference on Intelligent Computer Mathematics*, pp. 287–288, Springer, 2011.

[44] A. Armando and D. Zini, "Towards Interoperable Mechanized Reasoning Systems: the Logic Broker Architecture," in *AI*IA-TABOO Workshop: From Objects to Agents: Evolutionary Trends of Software Systems*, pp. 70–75, 2000.

[45] A. Adams, M. Dunstan, H. Gottliebsen, T. Kelsey, U. Martin, and S. Owre, "Computer algebra meets automated theorem proving: Integrating Maple and PVS," in *International Conference on Theorem Proving in Higher Order Logics*, pp. 27–42, Springer, 2001.

[46] J. Harrison and L. Théry, "A skeptic's approach to combining HOL and Maple," *Journal of Automated Reasoning*, vol. 21, no. 3, pp. 279–294, 1998.

[47] C. Ballarin, K. Homann, and J. Calmet, "Theorems and algorithms: An interface between Isabelle and Maple," in *Proceedings of the International Symposium on Symbolic and Algebraic Computation*, pp. 150–157, ACM, 1995.

[48] P. Jackson, "Exploring abstract algebra in constructive type theory," in *International Conference on Automated Deduction*, pp. 590–604, Springer, 1994.

[49] J. Siekmann, C. Benzmüller, V. Brezhnev, L. Cheikhrouhou, A. Fiedler, A. Franke, H. Horacek, M. Kohlhase, A. Meier, E. Melis, *et al.*, "Proof development with OMEGA," in *International Conference on Automated Deduction*, pp. 144–149, Springer, 2002.

[50] C. Ballarin and L. C. Paulson, "A pragmatic approach to extending provers by computer algebra—with applications to coding theory," *Fundamenta Informaticae*, vol. 39, no. 1, 2, pp. 1–20, 1999.

[51] R. Y. Lewis, "An extensible ad hoc interface between Lean and Mathematica." *in preparation*, 2017.

[52] E. Yudkowsky and M. Herreshoff, "Tiling agents for self-modifying AI, and the Löbian obstacle," tech. rep., Machine Intelligence Research Institute, 2013.

[53] P. LaVictoire, "An Introduction to Löbs Theorem in MIRI Research," tech. rep., Machine Intelligence Research Institute, 2015.

[54] M. Barasz, P. Christiano, B. Fallenstein, M. Herreshoff, P. LaVictoire, and E. Yudkowsky, "Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic," *ArXiv e-prints*, Jan. 2014.

[55] B. Fallenstein and N. Soares, "Problems of self-reference in self-improving space-time embedded intelligence," in *International Conference on Artificial General Intelligence*, pp. 21–32, Springer, 2014.

[56] N. Soares and B. Fallenstein, "Toward Idealized Decision Theory," *ArXiv e-prints*, July 2015.

[57] S. Garrabrant, T. Benson-Tilsen, A. Critch, N. Soares, and J. Taylor, "Logical Induction," *ArXiv e-prints*, Sept. 2016.

[58] K. Beck, *Test Driven Development: By Example*. Addison Wesley, 2002.

[59] R. Osherove, *The Art of Unit Testing: with examples in C#*. Manning Publications, 2013.

[60] E. M. Maximilien and L. Williams, "Assessing test-driven development at IBM," in *Proceedings of the 25th International Conference on Software Engineering*, pp. 564–569, IEEE, 2003.

[61] H. Erdogmus, "On the effectiveness of test-first approach to programming," *IEEE Transactions on Software Engineering*, vol. 31, no. 1, 2005.

[62] G. P. Sarma, T. W. Jacobs, M. D. Watts, S. V. Ghayoomie, S. D. Larson, and R. C. Gerkin, "Unit testing, model validation, and biological simulation," *F1000Research*, vol. 5, 2016.

[63] C. Keller, *A Matter of Trust: Skeptical Communication Between Coq and External Provers*. PhD thesis, École Polytechnique, 2013.

[64] M. Armand, G. Faure, B. Grégoire, C. Keller, L. Théry, and B. Werner, "A modular integration of SAT/SMT solvers to Coq through proof witnesses," in *International Conference on Certified Programs and Proofs*, pp. 135–150, Springer, 2011.

[65] J. Harrison, "Towards self-verification of HOL Light," in *International Joint Conference on Automated Reasoning*, pp. 177–191, Springer, 2006.

[66] R. Pollack, "How to believe a machine-checked proof," *Twenty Five Years of Constructive Type Theory*, vol. 36, p. 205, 1998.

[67] G. Hardy and J. Hadamard, "The Psychology of Invention in the Mathematical Field," 1946.

[68] S. Dehaene, *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press, 2011.

[69] P. Drijvers and K. Gravemeijer, "Computer Algebra as an Instrument: Examples of Algebraic Schemes," in *The Didactical Challenge of Symbolic Calculators*, pp. 163–196, Springer, 2005.

[70] P. Drijvers, "Learning mathematics in a computer algebra environment: obstacles are opportunities," *Zentralblatt für Didaktik der Mathematik*, vol. 34, no. 5, pp. 221–228, 2002.

[71] G. Lakoff and R. Núñez, *Where mathematics come from: How the embodied mind brings mathematics into being*. Basic books, 2000.

[72] S. Wolfram, *An Elementary Introduction to the Wolfram Language*. Wolfram Media, 2015.

[73] L. C. Paulson, "The foundation of a generic theorem prover," *Journal of Automated Reasoning*, vol. 5, no. 3, pp. 363–397, 1989.

[74] L. C. Paulson, *Isabelle: A generic theorem prover*, vol. 828. Springer Science & Business Media, 1994.

[75] Y. Bertot and P. Castéran, *Interactive theorem proving and program development: Coq'Art: The Calculus of Inductive Constructions*. Springer Science & Business Media, 2013.

[76] L. de Moura, S. Kong, J. Avigad, F. Van Doorn, and J. von Raumer, "The Lean Theorem Prover," in *International Conference on Automated Deduction*, pp. 378–388, Springer, 2015.

The Technological Landscape Affecting Artificial General Intelligence and the Importance of Nanoscale Neural Probes

Daniel Eth
Department of Applied Physics, Yale University, New Haven, CT, USA
E-mail: Daniel.eth@yale.edu

Keywords: AGI, *de novo* AGI, neuromorphic AGI, Whole Brain Emulation, nanotechnology, AI safety

Received: August 31, 2017

In this paper, we contrast three major pathways to human level AI, also known as artificial general intelligence (AGI), and we investigate how safety considerations compare between the three. The first pathway is de novo AGI (dnAGI), AGI built from the ground up. The second is Neuromorphic AGI (NAGI), AGI based loosely on the principles of the human brain. And third is Whole Brain Emulation (WBE), AGI built by emulating a particular human brain, in silico. Bostrom has previously argued that NAGI is the least safe form of the three. NAGI would be messier than dnAGI and therefore harder to align to arbitrary values. Additionally, NAGI would not intrinsically possess safeguards found in the human brain – such as compassion – while WBE would. In this paper, we argue that getting WBE first would be preferable to getting dnAGI first. While the introduction of WBE would likely be followed by a later transition to the less-constrained and therefore more-powerful dnAGI, the creation of dnAGI would likely be less dangerous if accomplished by WBEs than if done simply by biological humans, for a variety of reasons. One major reason is that the higher intelligence and quicker speed of thinking in the WBEs compared to biological humans could increase the chances of traversing the path through dnAGI safely. We additionally investigate the major technological trends leading to these three types of AGI, and we find these trends to be: traditional AI research, computational hardware, nanotechnology research, nanoscale neural probes, and neuroscience. In particular, we find that WBE is unlikely to be achieved without nanoscale neural probes, since much of the information processing in the brain occurs on the subcellular level (i.e., the nanoscale). For this reason, we argue that nanoscale neural probes could improve safety by favoring WBE over NAGI.

Povzetek: Analizirane so tri poti za dosego splošne inteligenca tipa človeške inteligence, poleg tega so analizirane potencialne nevarnosti in problemi.

1 Introduction

Scientists disagree about when humanity will develop artificial intelligence that is at least as smart as humans in most or all facets of intelligence, with common estimates ranging throughout the 21st century [1]. There's little disagreement, however, that such so-called *artificial general intelligence* (AGI) will be transformative.

The human species has used its high intelligence to influence the world more than has any other species, and an even greater intelligence in AGI could potentially influence the world even further. Many scientists therefore expect the creation of AGI to be the single most impactful advent in human history [2].

Consequently, there has been an increase in research into how to align AGI with human values (so that this impact is for the better) [3]. Most of this research focuses on a hypothetical AGI that's programmed from the ground up (de novo), and this de novo AGI (dnAGI) is often considered as an extension of existing machine learning research or as some other abstract utility-maximizing agent (such as AIXI) [4][5].

While creating dnAGI is one potential path to AGI, there are other paths as well. Comparatively little research has been performed investigating the risks and benefits from various avenues, despite the fact that each avenue poses different challenges.

In this paper, we investigate the major technological landscape leading to AGI, and we assess which technological trends appear likelier to favor positive and negative outcomes.

2 Three major paths to AGI

The three major paths to AGI are dnAGI, Neuromorphic AGI (NAGI), and Whole Brain Emulation (WBE). For dnAGI, computer programmers conceive of algorithms that yield intelligence. For NAGI, the human brain is studied, and certain key features of the brain's architecture are appropriated, yielding an intelligence with some similarities to human brains. For WBE, the brain of a particular human is scanned, this scan is translated into a model, and the model is run on a computer – yielding an intelligence similar to that of the person whose brain was scanned (the human is said to have been "uploaded").

Other paths to AGI exist, but for this paper, we will consider just these three. In many cases, insights about other paths can be inferred from insights about these three. For example, one other path is simulating a generic human brain instead of a specific one, and this sits somewhere between NAGI and WBE.

In *Superintelligence*, Bostrom posits that NAGI is the most dangerous of the three [6]. The logic here is straightforward and sensible. NAGI is much more "messy" than dnAGI and thus would be harder to align sufficiently with human values. WBE, while perhaps even messier than NAGI, inherently contains safeguards that NAGI and dnAGI do not – such as compassion and other human values (to the extent that the human being uploaded holds "human values").

It is still an open question whether WBE or dnAGI is safest. Bostrom initially appears ambivalent about this topic, but later implies that he'd prefer dnAGI. His main argument is that dnAGI is ultimately the most powerful kind of AGI, so humankind must undergo a potentially dangerous transition with the development of dnAGI, even if WBE had been developed beforehand. If instead dnAGI is developed first, WBE may still be developed later. But because dnAGI ultimately will be more powerful, we'll only face a risk from the first transition [6].

This argument presupposes that an advanced form of dnAGI would be more powerful than an advanced form of WBE. The architecture of the human brain fundamentally places a constraint on the capabilities of WBE, and this constraint does not exist for dnAGI. Tweaks may allow either of these technologies to reach a higher level of intelligence than that of any biological human, but the upper limit is presumably higher for dnAGI than for WBE, and the path of improvements is likely steeper for the less constrained dnAGI.

While we think there is some merit to this argument, ultimately getting WBE first may still be preferable. We would be remiss not to consider, however, that pursuing WBE might lead to the particularly unfortunate outcome of NAGI being developed first. Having said that, it is argued in this piece that it would be good to accelerate the development of WBE, insofar as this can be done in a manner which accelerates the development of WBE relative to that of NAGI to a significant degree (such that the chances of achieving WBE first increase, the chances of achieving dnAGI first decrease, and the chances of achieving NAGI first either decrease or stay the same).

Consider just how hard it may be to align a dnAGI to human values. Not only would we have to figure out how to align dnAGI to arbitrary values, but we'd also have to figure out how to specify human values in a manner the dnAGI would understand. How do you explicitly specify values such as fairness or happiness? This is prima facie a Herculean task.

Compared to the task of aligning dnAGI to human values, safely traversing the path to WBE seems relatively easy. If built correctly, WBE would be generally safe – even in the absence of significant work on AI safety. For dnAGI, on the other hand, this is not necessarily the case, and even large efforts specifically focused on AI safety might fail. Additionally, there is reason to believe that mistakes in WBE wouldn't be as dangerous as mistakes in dnAGI. For WBE, minor mistakes may be tolerable, since the brain is resilient to small perturbations (there is no reason to expect such a safeguard in dnAGI). Also, screening WBE for safety would be much easier than screening dnAGI for safety, since we have the fields of psychology and psychiatry that may help us diagnose antisocial tendencies in WBE.

Even if a WBE was unsafe, that situation itself would be much less dire than an unsafe dnAGI. It seems unlikely that the initial arrival of WBE will mark an immediate artificial intelligence "takeover" – after all, we already have 7 billion beings with human level intelligence, and no one has been able to accomplish such a takeover. On the other hand, the first arrival of dnAGI might be vastly more capable than the smartest humans and might quickly take over. Either WBE or dnAGI could iteratively improve its own code, leading to an intelligence explosion. For dnAGI, this "intelligence takeoff" scenario (which must not be confused with the separate event of a "takeover," previously mentioned) could be very fast, while for WBE, it seems quite unlikely that the takeoff would be anywhere near as fast. The messiness of the brain's architecture and the constraints of the brain may limit how quickly the intelligence of a WBE could be improved. Since WBEs would have a harder time quickly taking over or cognitively taking off, humanity would likely be able to pull the brakes on a dangerous WBE in a way that we might not be able to with a rogue dnAGI.

Even if WBE is developed first, the subsequent shift from WBE to dnAGI would likely be a much bigger risk than the initial shift to WBE. Therefore, the major risk associated with the development of WBE is likely not from the shift to WBE itself, but in how WBE would affect the shift to dnAGI.

If we upload humans who are particularly intelligent and ethical, the resultant WBEs would be the very agents we would want to work on the problem of creating dnAGI safely. One reason to in general expect the uploading of humans that are at least relatively ethical is that more people would probably want to upload people significantly more ethical than themselves than would want to upload those significantly less ethical than themselves – especially given the stakes. By no means should we assume that only ethical people would ever be uploaded, as unethical people might have the means to get themselves uploaded. But – especially when WBE technology is new and there is likely to be much public debate about who should be uploaded and how they should be chosen – it is likely that on average WBEs would be more ethical than the general population as a whole.

With enough hardware, the WBEs would be able to think much faster than biological humans. Since modifications in WBEs would be much easier than modifications of biological humans, WBEs could self-modify more than could biological humans. Every time a modification was introduced, there would be a risk of value drift, but WBEs would presumably want to avoid

modifying themselves in ways to become the kind of agents that they didn't want to become. If the WBEs were taken from diverse cultures, a well-coordinated group of such WBEs would begin to embody what Yudkowsky has dubbed "coherent extrapolated volition" – an idea that AGI would be best to do what humanity wanted, if we "knew more, thought faster, were more the people we wished we were, had grown up farther together; … where our wishes cohere rather than interfere; [etc]."[7]

In *The Age of Em*, Hanson uses well-accepted, academic science (both hard science and social science) to predict the broad strokes of a potential future world dominated by WBEs [8]. His analysis is useful for our purposes of evaluating dnAGI created by WBEs, as we can consider how the world Hanson describes could affect dnAGI safety concerns. Hanson's analysis focuses on a period when WBE technology has advanced to the point that renting WBEs for labor is generally cheaper than paying biological humans for the same labor. Market forces cause the number of WBEs to increase rapidly, and WBEs perform almost all labor previously performed by biological humans [8]. Since the primary driving forces in this scenario for the creation of WBEs (either by uploading new humans or copying existing WBEs) are economic, WBEs are selected to be particularly profitable workers. Additionally, training for WBEs and mind "tweaks" would be selected for their effects on profitability [8]. In this scenario, previous considerations about WBEs being particularly ethical due to the desire to upload particularly ethical people may not hold; market forces would presumably present a much stronger selection effect.

Hanson's application of economics, sociology, and psychology (among other disciplines) leads to many conclusions about how WBEs might live [8]. Most of these conclusions aren't obviously related to the safety of subsequently developed dnAGI, but some are. Hanson argues that compared to most biological humans today, we should expect most WBEs to be smarter, more rational, more work-oriented, more mindful, more patient, and less mistake-prone [8]. All of these traits imply a lower chance of making a mistake in AI alignment. Additionally, Hanson argues that we have at least weak evidence to expect most WBEs (again, compared to biological humans today) to be better coordinated, more law-abiding, more trustworthy, and more expectant of and focused on preparing for big disasters [8]. These traits seem to imply WBEs would have a greater chance of successfully coordinating to prevent the creation of an unsafe dnAGI. On the other hand, the world Hanson describes is one that is much more economically competitive than our current world, which would perhaps increase the chance of a tighter race for the development of dnAGI [8]. This competition plausibly could lead firms to neglect important safeguards in dnAGI that they might consider to be luxuries they could not afford.

Implicit in Hanson's analysis is an assumption that WBE will be achieved long before dnAGI otherwise would have been achieved [8]. After WBE is developed, the cost for WBEs would need to fall below the cost of biological human labor for most jobs, and then economic equilibrium would need to more or less be established – all without dnAGI or NAGI being developed in the meantime (even with WBEs working towards creating these other forms of AGI).

If WBE is the first type of AGI created, it remains to be seen whether or not the economic conditions necessary for the Hansonian scenario will be realized before other forms of AGI are developed. Either way, shortly after the development of WBE, many WBEs will likely be significantly more competent than most biological humans – at least on many matters relevant to dnAGI safety.

Would a team of such WBEs be able to create a safe dnAGI? Possibly. But if such a highly competent team cannot, it's even less likely that biological humans could solve the problem.

Furthermore, such WBEs may better enable dnAGIs to be taught human values. Some proposals for specifying human values do not involve explicitly specifying the values, but instead involve allowing AI to learn the values from humans. For instance, AI could theoretically glean human values by observing humans, and then the AI could further be trained with feedback on its behavior provided by humans [9]. One limitation with this approach is that the amount of data and feedback that could be produced in any timeframe would be limited. More data and feedback could be produced with WBEs, since WBEs could be run to think much faster than biological humans (Hanson has estimated that WBEs might generally think about 1,000 times faster than biological humans, and perhaps sometimes even 1,000,000 times faster or more) [8].

Another idea has been to interconnect AI into our nervous systems. In this scheme, humans would more or less act as the limbic system for an AGI that would carry out our wishes [10]. This approach has several limitations, and again WBEs would be quite helpful. One major limitation is the difficulty in physically integrating AI with the processing of the brain. This would surely be much easier to do with a virtual brain than in a biological brain. Another difficulty is that such human-AI systems would be limited in speed by the human thinking, and other, separate AGI would surely be faster. Again, WBEs – operating much faster than biological humans – might be fast enough for this proposal to actually work.

While it's possible that NAGI might be developed after WBE (and before dnAGI) and that NAGI may also be more powerful than WBE, similar arguments to above apply for why developing WBE before NAGI would likely be safer than developing NAGI without first developing WBE. It must be acknowledged, however, that the existence of WBE might make NAGI easier due to their similarity. On the other hand, insofar as this would be a dangerous path to follow, smart WBEs might be able to avoid it. Additionally, even if NAGI could be achieved sooner than dnAGI, the fast speed of thinking in WBEs might decrease the cost (in time) of simply waiting for dnAGI without first developing NAGI.

Taking all of the above into consideration, plotting the expected safety of different types of AGI versus their similarity to a human brain yields a J-curve (Figure 1).

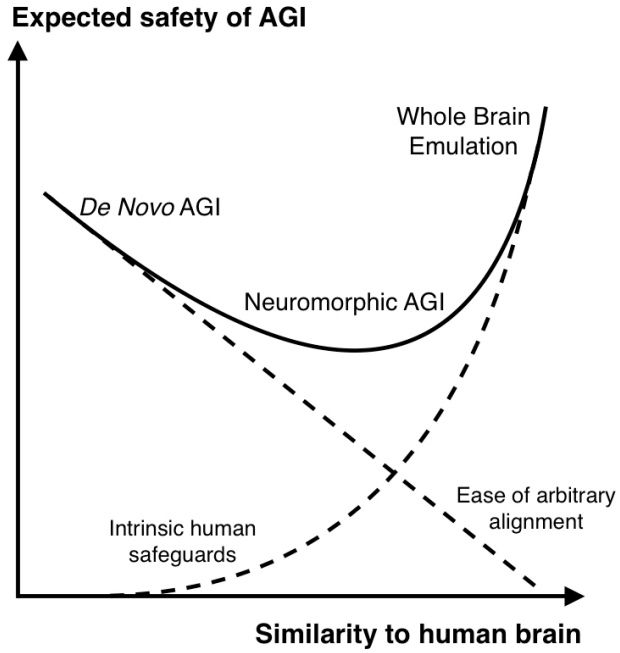


Figure 1: Expected safety of different types of AGI versus their similarity to the human brain. The downward sloping dotted line represents the fact that an AGI more similar to the human brain will tend to be messier and harder to align to arbitrary values. The upward sloping dotted curve represents the fact that an AGI especially similar to the human brain intrinsically contains human values. Taking these two effects into account reveals the solid curve to be a J-curve.

3 Technological landscape

Naively, we should work towards creating the type of AGI that would be safest if completed first. However, we must additionally consider interplay between different technologies. For example, while WBE may be safer than dnAGI, technology that progresses us towards WBE will often also progress us towards NAGI. This line of reasoning has led Bostrom to posit that *even if* we consider WBE to be safer than dnAGI, it *still* might be preferable to promote dnAGI so as to avoid NAGI [6].

This is reasonable as a general rule. When looking at specific technological trends, however, it is useful to consider the broader context of other relevant technological trends. In this section, we first will attempt to elucidate the major technological trends leading to each of the three major types of AGI. We will then combine these trends into a broader technological landscape and argue which trends are best to advance from an AI safety perspective.

3.1 *De Novo* AGI

The two major enabling technologies for dnAGI are computational hardware and what we will simply call "AI research," meaning certain types of software research (such as machine learning) that may be used in dnAGI (but not necessarily directly in WBE/NAGI). Much of current AI research doesn't progress us closer to dnAGI, and very little AI research is performed explicitly to direct us to AGI. In this paper, we're using the phrase "AI research" to mean any of this research that does progress us in the direction of dnAGI, whether or not the research is being performed for that purpose. Computational hardware and AI research, taken together and at an advanced enough level, seem necessary and sufficient for dnAGI. Interestingly, advances in both hardware and in AI research should lead to advances in the other one of the two. Improvements in hardware can lead to improvements in software for a variety of reasons, including allowing for more rapid testing of algorithms, and greater use of computationally heavy methods of algorithm design, such as genetic algorithms (which use Darwinian pressures to design and select algorithms according to certain criteria) [11]. Improved AI algorithms can be used to find superior computer chip designs. Hardware itself has a positive feedback loop, as greater computational capabilities are useful in powering the algorithms that help design chips.

Further upstream, nanotechnology research is a major enabler of improved hardware – especially as we reach the limits of silicon devices and other materials are needed to take over (options for such materials includes carbon nanotubes). Since one major subfield of nanotechnology is computational nanotechnology (using computers to advance nanotechnology, such as by simulating materials on the nanoscale), improved hardware and AI software would also aid nanotechnology research.

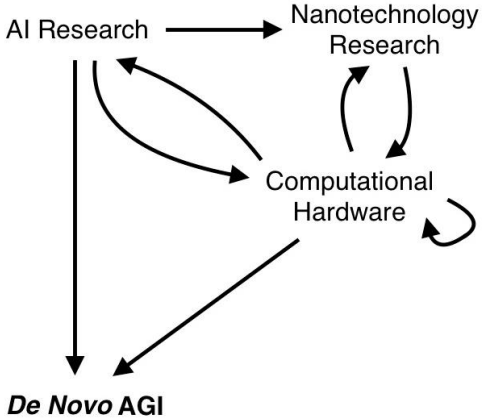


Figure 2: Technological pathway to *de novo* AGI. AI research and computational hardware are the main technological requirements.

3.2 Neuromorphic AGI

For NAGI, the major enabling technologies are hardware and neuroscience. With enough of the right kinds of neuroscientific knowledge to create algorithms of intelligence, and enough hardware to run such algorithms, NAGI could be achieved. Hardware was already discussed in the section on dnAGI. Since

computational neuroscience is a major aspect of neuroscience, both hardware improvements and AI research would also aid neuroscience (in much the way that they'd also aid nanotechnology). Nanotechnology could also provide much benefit for neuroscience through the creation of nanoscale neural probes. Such probes would have many benefits for neuroscience over existing brain scanning technologies – in particular, they could scan the brain with subcellular resolution, *in vivo*, and many could potentially be used in parallel to determine the architecture and function of neural circuits. Advances in neuroscience would additionally be useful for designing such probes.

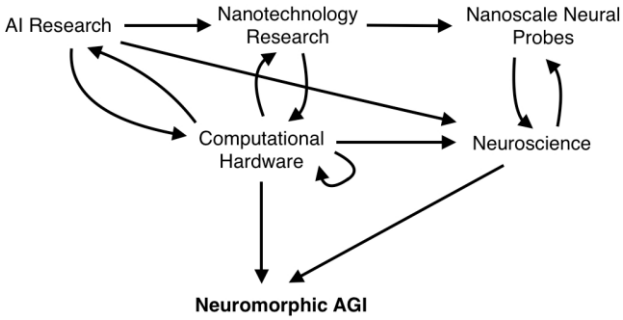


Figure 3: Technological pathway to Neuromorphic AGI. Neuroscience and computational hardware are the main technological requirements.

3.3 Whole Brain Emulation

The technological landscape of WBE looks quite similar to that of NAGI. Like NAGI, WBE would require computational hardware and neuroscientific knowledge. It should be noted, however, that the specific hardware and neuroscience requirements for WBE and NAGI may differ. WBE presumably requires more hardware, since WBE would be forced to simulate many details of brain function. In terms of neuroscience, NAGI would require greater conceptual understanding, while WBE would require greater understanding of details.

Another major difference between WBE and NAGI is the role of nanoscale neural probes. As we have argued elsewhere, it is unlikely WBE will be achieved without nanoscale neural probes [12]. In order to create a model of the human brain with enough fidelity to allow for WBE, we will arguably need the ability to study interactions within synapses, *in vivo*, and at large scale. Clearly, destructive brain scanning techniques (such as scanning electron microscopy) cannot achieve this alone, as these techniques destroy the brain and can therefore only be used to determine structure – not brain activity. Current large scale, nondestructive brain scanning techniques (such as MRI) don't appear capable of fulfilling this task either, as their resolution is too limited (and will likely run into harder limits such as any imposed by the skull). Single cell techniques (such as the patch clamp) can monitor single neurons or small groups of neurons for a few signals, but not large circuits of neurons for many types of chemicals. On the other hand, yet to be developed nanoscale neural probes would be able to fulfill all these tasks.

It should be noted that Sandberg and Bostrom have argued that WBE could be achieved without nanotechnology. They propose that brain architecture could be determined by automation of destructive scanning techniques similar to those that exists today (such as electron microscopy), using many of such automated machines in parallel. In order to model neuronal activity to the necessary precision, they suggest using a combination of this large-scale scanning and wet experiments (such as *in vitro* experiments on neurons) to create models, which can then be analyzed and used to guide further experiments, until the model is sufficiently refined [13]. We are personally very doubtful that such a scheme would allow for gathering the neuroscientific detail necessary for WBE. It is a well-known understatement to say that the brain's method of information processing is complicated, but what's not well appreciated is that this complexity in information processing doesn't just apply to how neurons are arranged – it includes many subcellular processes. Historically, scientists thought of the brain as consisting of a bunch of neuronal nodes that pass information simply in the form of "spikes" across passive synapses, yet we now know that reality is not so simple. In reality, neurons aren't simply nodes, but instead there is a large diversity of neuron types (with different behaviors), and computation is performed within the neuron cell bodies, within the axons, and within the dendrites [14]. Neurons don't communicate just through electrical signals either – around 10 common neurotransmitters and 200 uncommon neuromodulators are implicated in neuronal interaction, and neurons can even communicate without direct synaptic communication, such as via ephaptic coupling (nerve fiber coupling via local electric fields) and via chemical diffusion in the extracellular space [15][16]. Synapses themselves show a large diversity of types, and far from being simply conveyers of information, they play an active role in information processing [17]. In addition to neurons, glial cells (brain cells that outnumber neurons 10:1 but that have historically been ignored since they do not communicate via electrical impulses) can influence neurotransmission [18]. These findings have largely come as surprises to the scientific community, and it would be naïve to assume that we won't find any more surprises. The overall picture of the brain that we get from considering these findings is that much information processing occurs in the brain on the subcellular scale, which happens to be the nanoscale. In order to understand this information processing well enough to perform WBE, it is likely that nanotechnology will be essential. In biology, systems can act quite different *in vitro* versus *in vivo*. For a system as complicated as the brain, this would be particularly expected. Therefore, not only would brain activity likely need to be studied on the nanoscale, but it would likely need to be studied on the nanoscale *in vivo*. Nanoscale neural probes are the only foreseeable technology that could yield such information. Incidentally, several ideas for using nanotechnology to map the activity in the brain have been proposed, and the general idea of using nanotechnology to map activity in the brain is central to

the United State's multi-year, multi-billion dollar BRAIN Initiative [19][20].

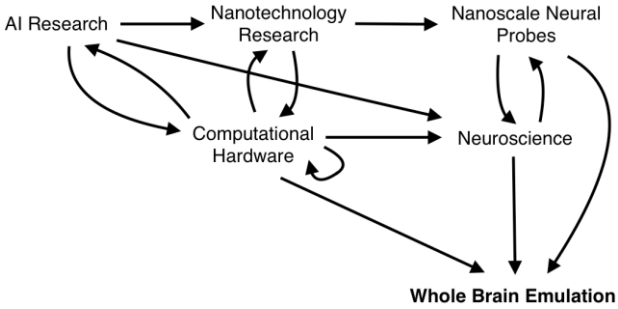


Figure 4: Technological pathway to Whole Brain Emulation. Computational hardware, neuroscience, and nanoscale neural probes are the major technological requirements. Note that the biggest difference between the pathway to Neuromorphic AGI and Whole Brain Emulation is the role of nanoscale neural probes in WBE.

While nanoscale neural probes are likely necessary for WBE, they are less likely to be necessary for NAGI. Since NAGI wouldn't need a one-to-one mapping with a particular brain, the specific details about the roles of each of these subcellular parts may not be needed – if certain general principles of how the brain processes information can be found via methods other than nanoscale neural probes, that might be good enough for NAGI. Even if there are gaps in the understanding of how the brain operates, it may be possible to create NAGI without understanding those gaps, by instead developing other algorithms that process information in a manner different from exactly how the brain does, yet that still accomplish the same general tasks.

3.4 The larger picture

Putting the aforementioned trends into a larger technological landscape yields a complicated and interconnected picture. Since our reasoning has included several simplifications (such as breaking AGI into only three distinct types), the real picture is undoubtedly more complicated. Accordingly, we must recognize that most implications are uncertain and open to revision upon further analysis. Having said that, we believe several implications can be produced.

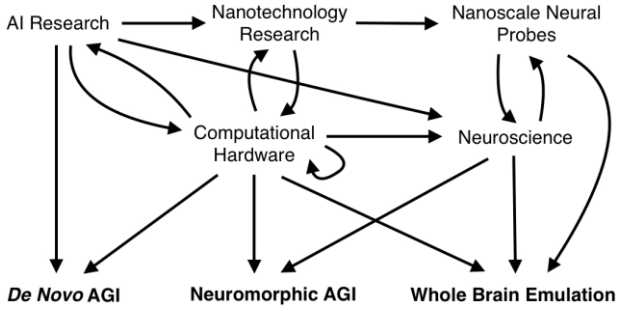


Figure 5: The technological landscape for the three main paths to AGI.

Computational hardware is the only major technological trend directly required for all three types of AGI, and it has a complicated relationship with AI safety. Larger computational hardware does have the potential benefit of favoring WBE over NAGI (compared to situations where smaller hardware may allow for NAGI but not WBE). On the other hand, there are a couple possible problems associated with computational hardware getting *too* large. First, such a situation could allow for a "hardware overhang" – where the massive computational resources available at the introduction of AGI enable early AGI to be particularly powerful or large in number. This situation may be more disruptive than if AGI were created in a context without such an overhang. Second, larger computational resources might allow for more dangerous methods of creating AGI. For example, it may be possible to "brute-force" an AGI without understanding it well. Alternatively, such large resources may allow for creating AGI through genetic algorithms. Since the use of genetic algorithms can lead to surprising results, even if the starting inputs for AI were relatively well aligned with human values, mutation and Darwinian pressures could cause the values to drift considerably astray.

AI research and neuroscience research similarly hold vague positions regarding safety. The more one favors dnAGI over WBE/NAGI (either by thinking WBE isn't much safer than dnAGI, or that if we pursue WBE/NAGI we will probably end up with NAGI) the more one should support AI research above neuroscience research (and the other way around if you disfavor dnAGI).

Development of both of these technologies requires caveats, however. For dnAGI, it is important that so-called AI safety research keeps pace with AI research, such that dnAGI isn't developed before we can align it with human values. For neuroscience, it may be bad if nanoscale neural probes lag significantly behind the neuroscience, as that state of affairs favors NAGI over WBE.

For each of these areas, it is more prudent to focus on accelerating the safeguard technology (AI safety research and nanoscale neural probes) instead of attempting to slow down the technology that poses a risk. For both of these areas, there are many more people working on the risky technology than on the safeguard technology, meaning that an individual can likely have a proportionately larger impact on development of the safeguard technology. Vested interests additionally imply that slowing down the risky technologies would be quite difficult. Advocating for the slowing of progress on such technologies could additionally cause a backlash, leading to actors in the field to dismiss all calls for safety as alarmist or neo-Luddite. Furthermore, even if calls for slowing down the risky technologies led certain safety conscious actors to refrain from pursuing such technologies, that would mean those left to pursue them would be less safety conscious (and in a worst case scenario, development of such technologies could be pushed underground).

Nanoscale neural probes require a bit of an elaboration, since their impact is more nuanced. Not only would they favor WBE over NAGI, but they also would favor NAGI over dnAGI. Even if one is generally skeptical of the WBE/NAGI pathway, these probes still

might provide promise. Since NAGI and WBE share many characteristics, it is reasonable to assume that they might be developed relatively close in time to each other, ignoring the impact that one form of AGI would have on creating another form. Since these two technologies are quite different from dnAGI, it is also reasonable to assume a greater chance that NAGI and dnAGI would be developed further apart from each other in time (again, ignoring the fact that one form of AGI could help create other forms). Therefore, even if advancements in nanoscale neural probes not only accelerate WBE as compared to NAGI by a certain amount of time (say, years), but also accelerate NAGI over dnAGI by a similar amount of time, the net effect may still be a decrease in the likelihood of NAGI. This is because if NAGI and WBE are in a closer race, it is more likely to tip the scale from NAGI to WBE than it is to tip the scale from dnAGI to NAGI in a less close race.

Nanoscale neural probes may further provide a particularly powerful means of human augmentation. Such augmentation wouldn't on its face favor any particular technological trend over any other one, but may well increase general human competence and decrease the chance of making a mistake in AI alignment. Furthermore, such probes may enable improved brain computer interfaces. This could aid proposals for making AI safe by having the AI act as an extension of humans.

4 Conclusion

In this piece, we examined safety concerns around the three major proposals for AGI: dnAGI, NAGI, and WBE. NAGI likely would be the least safe due to messiness and lack of other safeguards, and the inherent human safeguards in WBE would likely make it the safest. Even though with WBE a second transition to dnAGI would likely subsequently take place, we argued that a WBE-first path is still preferable (assuming such a path advanced WBE over NAGI to an extent that the chances of getting NAGI first did not increase), since WBE could aid a safer transition to dnAGI.

We also examined the major technological trends leading to these three types of AGI. While explicit AI safety research has been accepted as a means to ensure dnAGI is safe, we found that another path to increasing AI safety could be provided by the development of nanoscale neural probes, which would favor WBE over NAGI.

Of all the trends we've examined, nanotechnology research, and relatedly nanoscale neural probes, has traditionally been the most neglected by the AI safety community. This makes sense, given the fact that nanotechnology research is further removed from AI research than any of the other trends listed, and because it isn't obviously related to AI, in contrast with AI research, hardware, and neuroscience.

Since nanotechnology holds many implications for AGI, further research should not ignore the implications that nanotechnology may hold.

5 References

[1] Müller, V. C. and Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. *Fundamental Issues of Artificial Intelligence*, Synthese Library; Berlin: Springer, 553-571.

[2] Hawking, S. *et al*. (2014). Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence – but are we taking AI seriously enough?' *The Independent*.

[3] Farquhar, S. (2017). Changes in funding in the AI safety field. *The Center for Effective Altruism*. https://www.centreforeffectivealtruism.org/blog/changes-in-funding-in-the-ai-safety-field/

[4] Amodei, D., Olah, C. *et al*. (2016). Concrete Problems in AI Safety. *ArXiv:1606.06565 [cs.AI]*.

[5] Hutter, M. (2000). A Theory of Universal Artificial Intelligence based on Algorithmic Complexity. *ArXiv:cs/0004001 [cs.AI]*

[6] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

[7] Yudkowsky, El. (2004). Coherent Extrapolated Volition. *The Singularity Institute*, San Francisco, CA. https://intelligence.org/files/CEV.pdf

[8] Hanson, R. (2016). *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. Oxford University Press.

[9] Amodei, D. Interviewed by Wiblin, R. (2017). Podcast: How to train for a job developing AI at OpenAI or DeepMind. *80,000 Hours.* https://80000hours.org/2017/07/podcast-the-world-needs-ai-researchers-heres-how-to-become-one/

[10] Urban, T. (2017). Neuralink and the Brain's Magical Future. *Wait But Why*. https://waitbutwhy.com/2017/04/neuralink.html

[11] Shulman, C. and Sandberg, A. (2010). Implications of a Software-Limited Singularity. *Machine Intelligence Research Institute*, Berkeley, CA. https://intelligence.org/files/SoftwareLimited.pdf

[12] Eth, D. *et al*. (2013). The prospects of whole brain emulation within the next half-century. *Journal of Artificial General Intelligence*, 4(3), 130-152.

[13] Sandberg, A. and Bostrom, N. (2008). Whole Brain Emulation: A Roadmap. *Future of Humanity Institute*, Oxford University, Technical Report #2008-3.

[14] Sidiropoulou, K. *et al*. (2006). Inside the brain of a neuron. *EMBO Reports*, 7(9) ,886-892.

[15] Anastassious, C. A., Perin, R. *et al*. (2011). Ephaptic coupling of cortical neurons. *Nature Neuroscience*, 14(2), 217-223.

[16] Vizi, ES, *et al*. (2010). Non-synaptic receptors and transporters involved in brain functions and targets of drug treatment. *British Journal of Pharmacology*, 160, 785-809.

[17] O'Rourke, N. A. *et al*. (2012). Deep molecular diversity of mammalian synapses: why it matters and how to measure it. *Nature Reviews: Neuroscience*, 13, 365-379.

[18] Fields, R. D. *et al*. (2014). Glial Biology in Learning and Cognition. *The Neuroscientist*, 20(5), 426-431.
[19] Marblestone, A. H., Zamft, B. M. *et al*. (2013). Physical principles for scalable neural recording. *Frontiers in Computational Neuroscience*, 7(137), 1-34.
[20] Editorial. (2014). Brain activity. *Nature Nanotechnology*, 9, 85.

M-learning Programming Platform: Evaluation in Elementary Schools

Efthimios Alepis and Christos Troussas
Department of Informatics, University of Piraeus, Piraeus, Greece
E-mail:talepis@unipi.gr, ctrouss@unipi.gr

Keywords: affective interaction, evaluation, intelligent tutoring system, mobile learning, programming

Received: January 19, 2017

Mobile learning has invaded the everyday life and therefore evaluation methodology and reporting should be used so that mobile technologies are tested out in a variety of learning contexts. To this direction, a mobile learning platform, named m-AFOL, was designed for elementary school students so that they are taught basic programming principles. The platform is a sophisticated learning tool that incorporates affective interaction among learners and tutoring agents in order to maximize educational gains. The paper presents a two-level evaluation study of m-AFOL, estimating its effectiveness and acceptance rate. The evaluation concludes that the mobile facilities of the resulting intelligent tutoring system are highly acceptable (over 80%) from young learners, being keen on using mobile devices while learning. Concluding. the mobile platform is evaluated in comparison with a desktop version of resulting system.

Povzetek: Razvita je bila inovativna mobilna platforma za učenje, imenova m-AFOL.

1 Introduction

As innovation proceeds with its quickly developing pace, individuals are increasingly involved in technological issues, including both software and hardware. Hence in recent years, the rapid development of high and new technology has opened new horizons in computer-assisted instruction (Troussas et al., 2013). At the same time, there has been an increasing focus on mobile software applications as a result of the rapid development of mobile networks (Troussas et al., 2014 a). Particularly in learning, individuals of all ages use such technology to bolster their knowledge through instruction. The ever increasing mobile population can assist mobile learning which focuses on the mobility of learners and instructors, interacting with portable technologies (Troussas et al., 2014 b). In our days, computer science is being taught in schools, even in lower school grades (Virvou et al. 2013). Young people figure out how to utilize computers, basic algorithmic principles and even pseudo code. In any case, most computer programming languages are far complicated to be taught as they require knowledge from many domains as prerequisites. Moreover, programming used to be quite incomprehensible in its infancy and even today it continues evolving to more "natural" languages for human beings. A characteristic example in this direction is the well-known programming paradigm of object oriented programming (Shieh et al. 1996), (Pastor et al. 2001).

(Abdul et al. 2013) investigated how children view technology according to their perspectives. This study concludes that children actually want technology which is ubiquitous, wearable, natural in interaction and child-centered. Towards this direction, an interesting study is of McKenney & Voogt (2010), who have examined 167 young children access, perceptions and use of technology within and outside of school settings. Their findings, among other, include the fact that children's attitudes toward computers are positive and also inform the debate on the desirability of young children's exposure to computers at home as well as in educational settings. The research presented in (Jackson et al. 2012) examined relationships between children's information technology (IT) use and their creativity, including 491 12-year-old participants. Regardless of race or gender, analyzing the participants' results indicated that using technological means and specifically videogame playing was associated with greater creativity. Over the last decade, parents from all over the world have been positively surprised, watching their young children use and "consume" technology related to their learning. Even very young children have shown comfort and confidence in using software and they can follow pictorial directions and use situational and visual cues to understand and think about their activities (Clements & Nastasi, 1993).

Teaching programming to young children, even in elementary schools is becoming quite popular over the last years. The term "programming" was probably unknown to the majority of people worldwide a few decades ago. However, as technology becomes an integral part of our everyday life, "programming" provides us, in a sense, with a way to communicate with computers. Younger children have a flair of learning easier foreign languages and this may also be the case with programming languages. The authors of (Fessakis et al. 2013) present a case study concerning the dimensions of problem solving using computer programming by 5-6 years old kindergarten children. The authors' research evidence supports the

view that children enjoyed the engaging learning activities and also had opportunities to develop mathematical concepts, problem solving and social skills. Kordaki (Kordaki, 2010) has presented and also evaluated a computer-based problem-solving environment named LECGO, designed for the learning of computer programming using Turbo C by beginners. The resulting learning environment has been evaluated by 12th grade students. The authors' findings from the evaluation study indicated that students gain better results within LECGO than in both the paper and pencil environment and the typical programming environment of Turbo C, while performing similar activities. Another interesting and quite relevant study is that of (Werner et al. 2012). In this paper, the authors describe a semester-long game-programming course where 325 middle school students used Alice, a programming environment. Their results of the analysis of 231 programming games show that many games exhibit successful uses of high level computer science concepts.

However, it is inferential that modern programming languages are not "designed" to be used by children, since their principal purpose was to be handled by grown-up programmers in order to produce effective and robust software applications. This comes in contrast with the aforementioned realization that children have a very efficient way of understanding and "communicating" with all kinds of technological means. In view of the above, the authors of this paper have proposed and designed a new programming language, incorporated in a mobile platform named m-AFOL (Alepis & Virvou, 2014). This system is targeted to children as end users and is highly user friendly and interactive since it is able to handle affect through a sophisticated tutoring agent module. This paper focuses on the evaluation of this platform in order to test its effectiveness as an educational tool and also to test its acceptance from elementary school students.

The remaining sections of this paper are organized as follows. In section 2 the authors present briefly an overview of the mobile learning platform with corresponding user interaction paradigms. In section 3 the settings of the evaluation study are shown, accompanied with its preliminary results. Section 4 shows a discussion about the significance and the importance of the evaluation study's findings. Finally, in section 5, the conclusions of this paper are presented.

2 Overview of the mobile learning platform

"Logo" programming language was introduced early in 1967 (Frazier, 1967). The programming language developers' objective was to take the best practices and ideas from computer science and computer programming and produce a user friendly interface that would be suitable for the education of young children. Logo has been used mainly in the past as a teaching language for children but its list handling facilities made it also useful for producing useful scripts. A detailed review on the "Logo" programming language from its early stages can be found in (Feurzeig, 2010).

Modern programming languages try to provide as much user-friendliness as possible while retaining their full programming functionality. Learning a programming language is a complex cognitive process, while it is argued that how people feel may play an important role on their cognitive processes as well (Goleman, 1981). At the same time, there is a growing number of researchers who acknowledge that affect has been overlooked by the computer community (Picard & Klein, 2002). Perhaps a remedy towards the problem of effectively teaching children through intelligent tutoring systems and educational applications may be found in rendering computer assisted and mobile-learning systems more human-like and thus more affective. Hence, the incorporation of emotion recognition modules as well as the interaction with animated tutoring agents in a tutoring system can be quite useful and profitable (Elliott et al., 1999). The presence of animated, speaking agents has already been considered beneficial for educational software (Johnson et al., 2000).

However, after a thorough investigation in the related scientific literature we come up with the conclusions that there is a shortage of educational systems that incorporate multi-modal emotion recognition, while we did not find any existing programming languages that incorporate emotion recognition and/or emotion generation modules in mobile devices. A relevant work in the personal computer dimension is that of Kahn (Kahn, 1996), where an animated programming environment for children is described, named ToonTalk.

The resulting platform named m-AFOL is the acronym for "mobile-Affective Object Oriented Logo Language", as a modernization of the past "Logo" programming language additionally including affective interaction and interaction with mobile portable devices. In the implementation of the m-AFOL language, the prevailing programming dimension is that of object oriented programming (Alepis & Virvou, 2014). Through the language's object-oriented architecture, children get familiar with the object oriented paradigm. Hence, an m-AFOL program may be viewed as a set of interacting objects, with their attributes and their methods, as opposed to the conventional Logo model, in which a program is seen as a list of tasks. Furthermore, the initial goal to create a programming language enjoyable by children is further improved through the m-AFOL language, by a highly user-friendly user interface, designed for affective interaction between high-tech mobile phones (incorporating their own Operating System) and young children.

A major concept in the presented educational platform is the mobile interaction. However, this kind of interaction has some prerequisites, regarding mobile hardware and software specification. More specifically, a smartphone with a multi-touch screen 3.7 inches or higher is required, while at this moment the application is available for the Android OS (Android version 4.0 and higher). Internet wireless connection is also required, since each mobile device acts as a client to a main web server who is responsible for handling and manipulating user models

and also for the system's affective interaction capability. This interaction is illustrated in figure 1.

Young students may use the educational application from a smartphone, having the opportunity to learn programming through m-learning courses. The theory is presented in graphical form while an animated tutoring agent is optionally but desirably present and may alternatively read the theory out loud using a speech

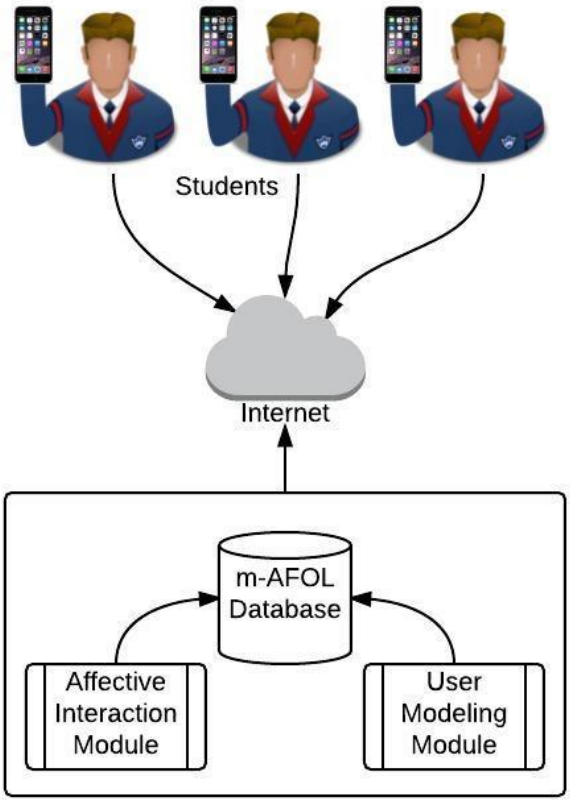


Figure 1: m-AFOL Architecture.

engine. Students are prompted to write simple programming commands producing lines and basic shapes as first steps towards getting familiar with the programming environment. More advanced level students are given graphical illustrations as examples and are prompted to write complete and sometimes quite complicated programs in the m-AFOL language in order to produce drawings, shapes and particular graphical illustrations. The mobile application is installed locally in each student's smartphone device, while an active internet connection is required for the communication with the main server. As it is already mentioned, the resulting application is targeted to the Android OS platform, while there are future plans for the integration of the resulting ITS to other mobile operation systems as well. Examples of interaction with the m-AFOL programming platform are illustrated in figures 2 and 3. It is noteworthy that in these examples the tutoring agent would also try to sympathize with the aims of each student trying to elicit human emotions. More specifically, figure 2 illustrates the programming mobile frame, where students may write programming commands. Special buttons help them store or load predefined "objects". Observing the m-AFOL's programming code, its Object Oriented "nature" is easily noticeable. Figure 3 illustrates the output of a student's programming code that results in the creation of a "sun" object and a "house" object. The animated agent (of the form of a cartoon computer) is congratulating the student for successfully completing a programming exercise.

3 Evaluation study

In the evaluation study, 40 elementary school students from two different classrooms participated (20 students from each class). The two classes were from the same

Figure 2: Programming in the m-AFOL mobile environment.

Figure 3: Programming result and interaction with the tutoring agent.

school and of the same grade. The school, that was chosen, is a public school, located in Athens, the capital city of the country. Hence, the school can be seen as a representative sample, since it adequately replicates the larger statistical population in terms of students' characteristics. School teachers also provided very valuable help in the whole evaluation study since they also participated both in the presentation of the ITS to the students and also provided assistance to their students while they interacted with the educational platform. The first class evaluated m-AFOL through mobile devices that were given to all students, while the second class evaluated the educational system through its conventional desktop interface that functions through personal computers. This division was very crucial in order to compare the mobile platforms effectiveness with the traditional way of tutoring through computers. As a result, both classes had given the appropriate hardware and as a next step they were given a brief presentation on how to use the educational platform. Consequently, each class had enough time to spend interacting and completing a whole lesson with m-AFOL, while all students had a small break in the middle of the session. After the completion of their interaction (class A with mobile devices and class B with personal computers), all students were given questionnaires to complete with guidance from the evaluators and also their teachers.

The evaluation study was conducted with the use of self-supplemented scale questionnaires incorporating closed questions for the students. For our research, we have used 28 questions regarding students:

- three (3) exploratory questions
- eight (8) questions regarding navigation in the platform
- six (6) questions regarding the user interface
- five (5) questions regarding hardware and software quality
- four (6) questions regarding evaluation of learning

It was observed that students became familiar easily and very quickly with the educational software, its features and its functionalities. Their interest was undiminished during the whole period of their interaction with the educational application.

Table 1 summarizes a basic set of questions that were asked to the students after their interaction with the application. These questions follow a 1 to 10 ranking (lower is negative, higher is positive) model.

Finally, figure 6 illustrates the statistical significance of questions 2, 7, 8 and 12. For the Null hypothesis:

"There is no difference between the two groups of students" the t-Test rejects the hypothesis for the fore mentioned 4 questions, while the remaining 8 questions seem to not have statistical differences between the two groups of students.

4 Discussion on evaluation results

The desktop version of m-AFOL, named AFOL has been presented in (Alepis, 2011). Providing a mobile interface for the educational platform was a great challenge since integrating all the system's functionalities in a mobile OS was a quite demanding task. This operation included supporting user modelling and also affective interaction in the new mobile interface. Both for building user models and also for the reasoning mechanisms used in affective interaction, sophisticated techniques including machine learning (Support Vector Machines) and stereotypic reasoning modules had to be used. Both the amount of data and also the processing mechanisms required hardware specifications that far exceed the hardware of the state of the art in modern smartphones. To address this problem, the authors of this paper have re-designed the system's architecture in order to function through a client-server web-service architecture. Thus, all "heavy" and resource demanding operations were processed by the main server, while the mobile devices were handled as user interfaces for the educational interaction, under the precondition of a stable internet connection between clients and the server.

N	Questions
1	Rate the application's user interface (1-10)
2	Rate your learning experience (1-10)
3	Did you like the interaction with the educational agent (1-10)?
4	Did the agent respond to your questions (1-10)?
5	Did the agent respond to your feelings/mood (1-10)?
6	Could you rate the affective interaction (1-10)?
7	Would you like to use this platform in your school (1-10)?
8	Did you find the tutoring system simple in use (1-10)?
9	Rate the overall quality of the ITS (Device and Software) (1-10)
10	Did you find the ITS helpful for your lesson (1-10)?
11	Would you suggest the ITS to your friends to use it (1-10)?
12	Rate the easiness in handling the device your were given (1-10)

Table 1: Sample of questions of the evaluation study.

It was expected that younger students with an inherent tend towards new technology would welcome mobile learning adapted to their needs, supporting their learning. The findings of this preliminary study are rewarding the authors' attempts towards moving education to the fast growing field of mobile computing and specifically to mobile intelligent tutoring systems.

Analyzing the results of the evaluation study there is considerable evidence that new mobile technology is quite welcome from young learners and could be incorporated in schools supporting the educational process.

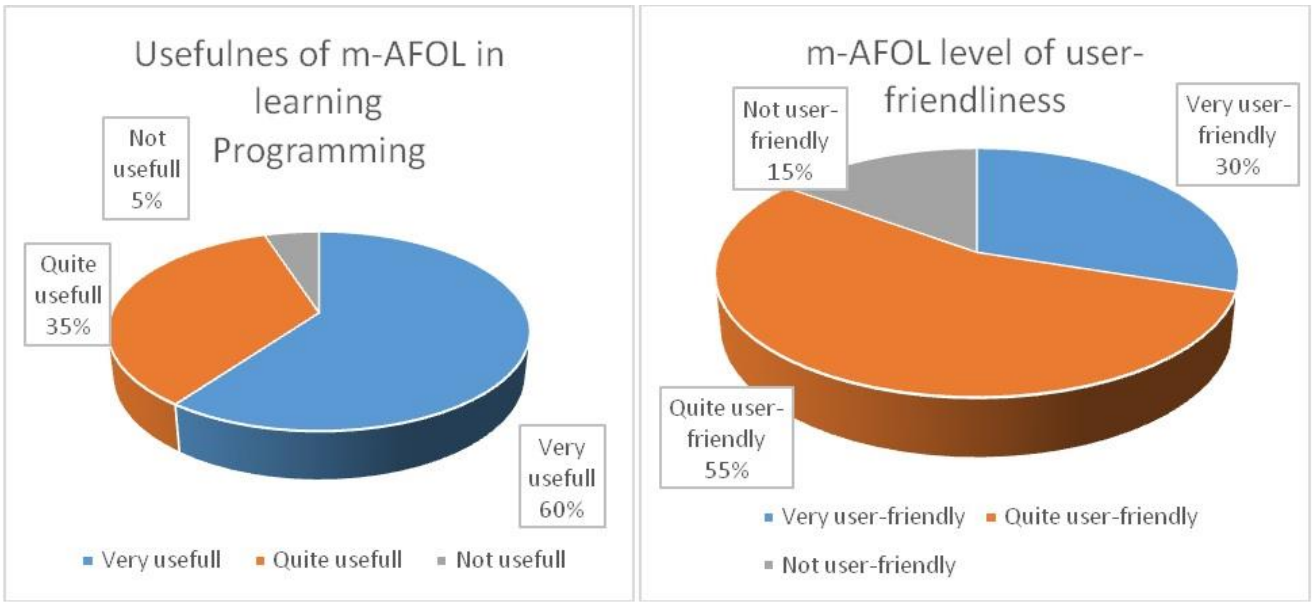


Figure 4: Usefulness and user-friendliness for the m-AFOL platform.

	1	2	3	4	5	6	7	8	9	10	11	12
Class A (m-AFOL)	8,4	8,5	8,4	8,1	7,5	7,7	8,7	7,4	7,7	7,9	8,5	6,9
Class B (AFOL)	8,1	7,3	7,8	8,2	7,4	7,5	7,5	8,4	8,1	8	7,8	8,5

Figure 5: Comparative ranking between the two instances of the educational platform.

The evaluation results are illustrated in figures 4 and 5 respectively. More specifically, figure 4 illustrates the usefulness and user-friendliness of m-AFOL, while figure 5 provides a comparative illustration between the mobile and the desktop version of the intelligent tutoring system, concerning the questions of table 1.

Figure 7 illustrates the whole procedure from the use of both applications (mobile and desktop) until the experimental results. However, there are design and user interface issues that should be addressed carefully, since mobile devices do not seem to have all the advantages a desktop computer could provide. For example in "easiness in use" of the system, the desktop version of the platform was quite prevailing. The same is the case of navigation where the traditional keyboard and mouse seem quicker options in human-computer interaction. However, mobile applications are far from reaching maturity and their domain is enriched on a daily basis. Perhaps, designing applications that do not only extend desktop versions but rather utilize all mobile functionalities could produce

t-Test: Two-Sample Assuming Equal Variances					
Question 2			**Question 7**		
	Variable 1	*Variable 2*		*Variable 1*	*Variable 2*
Mean	8.5	7.3	**Mean**	8.7	7.5
Variance	2.055556	2.233333	**Variance**	0.9	3.388889
Observations	20	20	**Observations**	20	20
Pooled Variance	2.144444		**Pooled Variance**	2.144444	
Hypothesized Mean Difference	0		**Hypothesized Mean Difference**	0	
df	18		**df**	18	
t Stat	1.832352		**t Stat**	1.832352	
P(T<=t) one-tail	0.041748		**P(T<=t) one-tail**	0.041748	
t Critical one-tail	1.734064		**t Critical one-tail**	1.734064	
P(T<=t) two-tail	0.083497		**P(T<=t) two-tail**	0.083497	
t Critical two-tail	2.100922		**t Critical two-tail**	2.100922	
Question 8			**Question 12**		
	Variable 1	*Variable 2*		*Variable 1*	*Variable 2*
Mean	7.4	8.4	**Mean**	6.9	8.5
Variance	0.711111	0.488889	**Variance**	3.211111	1.388889
Observations	20	20	**Observations**	20	20
Pooled Variance	0.6		**Pooled Variance**	2.3	
Hypothesized Mean Difference	0		**Hypothesized Mean Difference**	0	
df	18		**df**	18	
t Stat	-2.88675		**t Stat**	-2.35907	
P(T<=t) one-tail	0.004911		**P(T<=t) one-tail**	0.014912	
t Critical one-tail	1.734064		**t Critical one-tail**	1.734064	
P(T<=t) two-tail	0.009822		**P(T<=t) two-tail**	0.029824	
t Critical two-tail	2.100922		**t Critical two-tail**	2.100922	

Figure 6: Statistical significance in a student's t-Test.

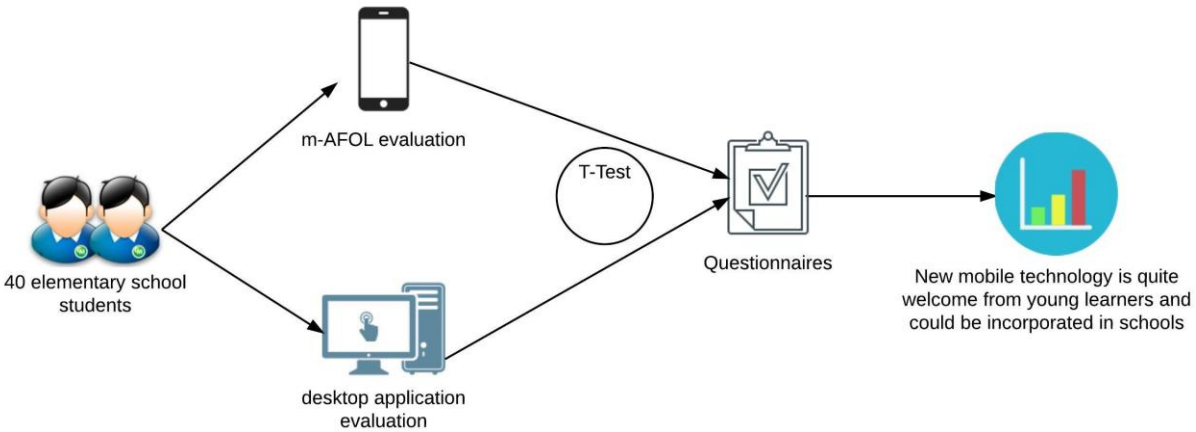


Figure 7: Experimental procedure.

more robust and sophisticated mobile systems. Modern mobile learning applications combined with their inherent portability and independence of time and place could be a great advance towards better and more natural computer assisted learning.

5 Conclusions and future work

In this paper, a mobile educational application for programming courses targeted to elementary school students has been presented and evaluated. The aim of this evaluation was twofold. First, we tried to measure the effectiveness of m-AFOL as an educational tool in a quite complicated domain of knowledge for young children (that of Object Oriented Programming). Secondly, we tried to compare the mobile platform with the same framework in a desktop computer interface. Analyzing the system's evaluation results gave us strong evidence that both students and also their instructors appreciated its functionalities and also approved it as an educational tool.

Mobile content in learning courses has been found quite friendly and attractive to younger students and also profitable for all potential learners.

Summarizing, the objectives of this paper is the presentation of the design of a mobile learning platform, named m-AFOL, for elementary school students so that they are taught basic programming principles and its evaluation. Furthermore, the primary goal of this study was the evaluation of the mobile programming platform in younger ages, namely children in elementary schools, given that students of this age are prone to new technologies. A contribution of the paper are the architecture of this platform being a sophisticated learning tool and incorporating affective interaction among learners and tutoring agents in order to maximize educational gains. Another contribution is that the paper presents a two-level evaluation study of m-AFOL, estimating its effectiveness and acceptance rate. The evaluation concludes that the mobile facilities of the resulting intelligent tutoring system are highly acceptable (over 80%) from young learners, being keen on using mobile devices while learning. Also, a quite interesting finding is that the majority of young children possessed mobile phones (2016-2017 findings) and know to use them.

Limitations and shortcomings can be the following. First of all, the population of the experimental group could include more than 40 elementary school students. Moreover, the evaluation stage may not let students interacting freely with the system and their behavior may not be the actual one. Finally, the mobile programming platform is targeted to the Android mobile operating system, given that it holds the greatest share of the global smartphone market; however, the mobile programming platform could be also targeted to the IOS mobile operating system and tablets or even smart televisions.

It is in our future plans to extend the system's functionalities by incorporating an authoring module that will give teachers the opportunity to edit the lessons remotely and also monitor the progress of their students. In this way, all users will further benefit from time and place independence during interacting with m-AFOL. Furthermore, our future plans include the evaluation of this mobile authoring tool; as such, the mobile application will be fully functional.

References

[1] Alepis, E., Virvou, M. (2014). *Object-Oriented User Interfaces for Personalized Mobile Learning*, Intelligent Systems Reference Library Volume 64, Springer Berlin Heidelberg, pp. 109-125.

[2] Abdul Razak, F.H. , Salleh, K. , Azmi, N.H. (2013). *Children's technology: How do children want it?*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Volume 8237 LNCS, pp. 275-284.

[3] McKenney, S., Voogt, J. (2010). *Technology and young children: How 4-7 year olds perceive their own use of computers*, Computers in Human Behavior Volume 26, Issue 4, pp. 656-664.

[4] Jackson, L.A., Witt, E.A., Games, A.I. , Fitzgerald, H.E. , Von Eye, A. , Zhao, Y. (2012). *Information technology use and creativity: Findings from the children and technology project*, Computers in Human Behavior Volume 28, Issue 2, pp. 370-376.

[5] Clements, D., Nastasi, B. (1993). *Electronic media and early childhood education*, B. Spodek, editor, Handbook of research on the education of young children, Macmillan, New York, pp. 251–275.

[6] Fessakis, G., Gouli, E. , Mavroudi, E. (2013). Problem solving by 5-6 years old kindergarten children in a computer programming environment: A case study, Computers and Education Volume 63, pp. 87-97.

[7] Kordaki, M. (2010). A drawing and multi-representational computer environment for beginners' learning of programming using C: Design and pilot formative evaluation, Computers and Education Volume 54, Issue 1, pp. 69-87 (2010)

[8] Werner, L., Campe, S., Denner, J. (2012). *Children learning computer science concepts via Alice game-programming*, SIGCSE'12 - Proceedings of the 43rd ACM Technical Symposium on Computer Science Education, pp. 427-432.

[9] Virvou, M., Alepis, E., Mpalasis, K. (2013). *Evaluation of a multimedia educational tool for geography in elementary schools*, Proceedings of International Conference on Information Communication Technologies in Education, ICICTE, pp. 364-374.

[10] Shieh, C.-K., Mac, S.-C., Chang, T.-C., Lai, C.-M. (1996). *An object-oriented approach to develop software fault-tolerant mechanisms for parallel programming systems*, Journal of Systems and Software, Volume 32, Issue 3, pp. 215-225.

[11] Pastor, O., Gómez, J., Insfrán, E., Pelechano, V. (2001). The OO-Method approach for information systems modeling: From object-oriented conceptual modeling to automated programming, Information Systems, Vol 26, Issue 7, pp. 507-534.

[12] Frazier, F. (1967). *The logo system: Preliminary manual*, BBN Technical Report. Cambridge, MA: BBN Technologies.

[13] Feurzeig, W. (2010). *Toward a Culture of Creativity: A Personal Perspective on Logo's Early Years and Ongoing Potential*, International Journal of Computers for Mathematical Learning , Volume 15, Issue 3, pp 257-265.

[14] Goleman, D. (1995). *Emotional Intelligence*, Bantam Books, New York.

[15] Picard, R.,W., Klein, J. (2002). *Computers that Recognise and Respond to User Emotion: Theoretical and Practical Implications*, Interacting with Computers, Volume 14, Issue 2, pp. 141-169.

[16] Elliott, C., Rickel, J., Lester, J. (1999). Lifelike pedagogical agents and affective computing: an exploratory synthesis. In Wooldridge MJ, Veloso M (eds) Artificial intelligence today, LNCS 1600. Springer, Berlin, pp. 195–212.

[17] Johnson, W.L., Rickel, J., Lester, J. (2000). *Animated pedagogical agents: face-to-face interaction in interactive learning environments*, International Journal of Artificial Intelligence 11, pp. 47–78.

[18] Kahn, K. (1996). *ToonTalk - An Animated Programming Environment for Children*, Journal of Visual Languages and Computing, Vol. 7, No. 2, pp. 197-217.

[19] Alepis, E. (2011). *AFOL: Towards a new intelligent interactive programming language for children*, Springer, Smart Innovation, Systems and Technologies Volume 11, pp. 199-208.

[20] Troussas, C., Virvou, M., Alepis, E. (2014). Collaborative learning: Group interaction in an intelligent mobile-assisted multiple language learning system, Informatics in Education, 13 (2), 279-292. (a)

[21] Troussas, C., Alepis, E., Virvou, M. (2014). *Mobile authoring in a multiple language learning environment*, IISA 2014 - 5th International Conference on Information, Intelligence, Systems and Applications, art. no. 6878819, pp. 405-410. (b)

[22] Troussas, C., Virvou, M., Alepis, E. (2013). Comulang: Towards a collaborative e-learning system that supports student group modeling, SpringerPlus, vol. 2, no. 1, pp. 1-9.

A Watermarking Algorithm for Multiple Watermarks Protection Using RDWT-SVD and Compressive Sensing

Rohit Thanki and Vedvyas Dwivedi
Faculty of Technology and Engineering, C. U. Shah University, Wadhwan City, Gujarat, India
E-mail: rohitthanki9@gmail.com, vedvyasdwivediphd@gmail.com

Komal Borisagar
Associate Professor, E.C. Department, Atmiya Institute of Technology & Science, Rajkot, Gujarat, India
E-mail: krborisagar@aits.edu.in

Surekha Borra
K.S. Institute of Technology, Bangalore, India
E-mail: borrasurekha@gmail.com

Keywords: biometrics, color image watermarking, compressive sensing, Compressive Sensing (CS) measurements, Redundant Discrete Wavelet Transform (RDWT), Singular Value Decomposition (SVD)

Received: June 8, 2017

In this paper, a watermarking algorithm is proposed and analyzed using RDWT-SVD and Compressive Sensing for multiple watermarks protection. In this algorithm, the multiple watermarks are inserted into single host medium. Here three watermarks are converted into its CS measurements using compressive sensing procedure before embedding into host medium. The CS measurements of watermarks are generated using discrete cosine transform (DCT) and normal distribution matrix. The singular value of these CS measurements of multiple watermarks is inserted into the singular value of approximation wavelet coefficients of R channel, G channel and B channel of color host image to get watermarked color image. The experimental results show that this proposed algorithm is equally worked for all types of watermarks. This proposed algorithm also provides robustness against various watermarking attacks and performed better than existed algorithms in the literature.

Povzetek: Opisani so algoritmi za zaščito vodotiska.

1 Introduction

Nowadays, research on digital watermarking is not a new phenomenon. Because of various digital watermarking algorithms are discussed and proposed by various researchers in last fifteen years [1–22]. These algorithms are designed for various types of information such as digital image, digital video, digital audio, and text. The digital watermarking algorithms can be classified into various categories based on processing domain, type of host medium, type of watermark information and type of application. The digital watermarking algorithms based on processing domain are divided into the spatial domain, transform domain and hybrid domain. In spatial domain watermarking, pixel information of host medium is modified according to watermark information. In transform domain watermarking, frequency coefficients of host medium are modified according to watermark information. The signal processing transforms such as fast Fourier transform (FFT), discrete cosine transform (DCT), discrete wavelet transform (DWT), redundant discrete wavelet transform (RDWT) and singular value decomposition (SVD) are used in transform domain based watermarking algorithms. In hybrid domain watermarking, hybrid coefficients (which is a combination of two or more transform coefficients) of host medium is modified according to watermark information. The limitation of spatial domain watermarking algorithms is that there are not secure against any manipulations [1, 21]. The transform domain watermarking algorithms have overcome the limitation of spatial domain watermarking algorithms but have less payload capacity [21]. The hybrid domain watermarking algorithms are performed better than spatial and transform domain algorithms.

The digital watermarking algorithms based on the type of host medium are divided into image watermarking, video watermarking, audio watermarking and text watermarking [21]. The digital watermarking algorithms based on the type of watermark information are divided into image watermarking and biometric watermarking [21, 23]. In image watermarking, standard image or image of the logo, the text is taken as watermark information and inserted into host medium. In biometric watermarking, biometric such as fingerprint, face, iris, speech, and signature is taken as watermark information and inserted into host medium. The digital watermarking algorithms based on the type of application are divided into robust watermarking and fragile watermarking. The robust watermarking is providing

protection against any manipulation and used for copyright protection of information [2-22]. The fragile watermarking is not providing protection but providing authentication against any manipulation and used for copyright authentication of information [1, 23-24].

The watermarking algorithms mentioned in [1-24] are used single watermark information and inserted into host medium. Therefore, the strong need of watermarking algorithm is required which can be inserted multiple watermarks information into host medium. So many researchers are discussed and proposed various watermarking algorithms which can be inserted multiple watermarks information. This type watermarking algorithm is called as multiple watermarking algorithms [25]. The multiple watermarking algorithms are divided into three types such as composite, successive and segmented [25]. The composite based multiple watermarking algorithms are inserted combined multiple watermarks as a single watermark and then inserted into host medium. The successive based multiple watermarking algorithms are inserted one by one watermark into host medium. The segmented based multiple watermarking algorithms are inserted multiple watermarks into the specific slot of host medium.

Recently, there are a lot of new watermarking algorithms are proposed by researchers for the security of various multimedia data. S. Borra and her research team proposed a lossless watermarking technique based on visual cryptography and central limit theorem for the protection of high-resolution images and sensitive images [26, 27]. Surekha Borra has also proposed a watermarking technique based on visual secret sharing for image protection [28]. There are also various new watermarking algorithms are proposed by N. Dey and his research team for the security of various biometric data such as fingerprint, retina, ECG signal and EEG signal [29-32]. These algorithms are designed using various transforms such as DCT, DWT, and SVD. These algorithms are used various approaches such as spread spectrum, an edge detection algorithm for achieved better results and security to data. In 2016, researchers have introduced a watermarking algorithm for 3D images protection [33].

In this paper, a watermarking algorithm is proposed for the security of multiple watermarks using stationary wavelet transformation (SWT), signature value decomposition (SVD) and compressive sensing (CS) theory. The SWT is also known as redundant discrete wavelet transform (RDWT). This algorithm provides protection to multiple watermarks against various watermarking attacks. This algorithm can be used for security of multiple watermarks transferred over the non-secure communication channel. The CS theory provides security to watermark data before embedding into host medium. This step is introduced one additional security layer in conventional watermarking approach. This algorithm can be used for security of biometric data in the multimodal biometric system. Using this algorithm, a user can transfer any important multimedia data or biometric data over a non-secure channel or between two modules of the biometric system. This algorithm can provide copyright protection to multimedia data because multiple watermarks can be embedded into the host. Because imposter can not generate secure watermark data without information of orthogonal matrices U, V and embedding factor. U, V and embedding factor are used as a secret key in this algorithm.

The rest of paper is organized as follows: in section 2, related work to proposed algorithm is given. In section 3, mathematics and information on redundant discrete wavelet transform (RDWT); Singular Value Decomposition and CS theory are presented. Section 4 gives information on the implementation of proposed algorithm with multiple watermark insertion. The result and discussion for robustness and performance of proposed algorithm for different watermarks information against various watermarking attacks are given in section 5. Finally, the conclusion is given in section 6.

2 Related work

The review on watermarking algorithms mentioned in [25, 34–40] is related to proposed watermarking algorithm. Authors in [25] proposed a watermarking algorithm based on DWT and DCT for multiple biometric watermarks insertion. In this algorithm, first face information is inserted into host image to get face watermarked image. Then speech information is inserted into face watermarked image to get face-speech watermarked image. Finally, signature information is inserted visibly on DCT coefficients of face-speech watermarked image to generate multiple watermarks based watermarked image.

Authors in [34] proposed a watermarking algorithm based on visual cryptography for multiple watermarks insertion. In this algorithm, three watermarks information is inserted into Y component of the color image. Authors in [35] proposed a watermarking algorithm based on SWT and spread spectrum for EMG signal protection. Authors in [36] proposed a watermarking algorithm based on DWT for two watermarks insertion. In this algorithm, two watermarks information is inserted into LL subband and HH subband of host image to get watermarked image. This algorithm is robust against all type watermarking attacks.

Authors in [37] proposed a watermarking algorithm based on RDWT for biometric watermarks. In this algorithm, speech watermark information is divided into two portions and then inserted into wavelet coefficients of red channel and blue channel of the color face image. This algorithm provides robustness against watermarking attacks. Authors in [38] proposed correlation based and spread spectrum based watermarking algorithms for multiple watermarks insertion. In this algorithm, first watermark information is inserted into host image to get the first watermarked image. Then second watermark information is inserted into first watermarked image to generate multiple watermarks based watermarked image. These algorithms are spatial domain algorithms because here pixel information of host image is modified according to multiple watermarks.

Authors in [39] proposed RDWT and independent component analysis (ICA) based watermarking algorithm for multiple logo insertions. In this algorithm, multiple watermark logos are inserted into LH and HL subbands of the host image. This algorithm provides robustness against watermarking attacks. Authors in [40] proposed DWT and CS theory based watermarking algorithm for multiple biometric watermarks insertion. In this algorithm, CS measurements of multiple biometric watermarks are inserted into HH subband of various level of host biometric image. The payload capacity of watermarking algorithms mentioned in [34-40] is up to 50%.

After discussion on reviewed papers, it is cleared that most existed algorithms are based on successive based multiple watermarking with having less payload capacity. Also, the most of the existed algorithms are used grayscale host image for multiple watermarks insertion. These existed algorithms are also provided copyright protection to multiple watermarks but applied on one type of watermark information either image or biometrics. Thus, in this paper, a hybrid watermarking algorithm is proposed which focuses on high payload capacity. The motivation of the present work arises from developing a watermarking algorithm which inserted multiple watermarks information. In this paper, an algorithm is proposed which embeds multiple watermarks information in the red channel, green channel and blue channel of color host medium. The color host image is decomposed using redundant discrete wavelet transform (RDWT) and singular value decomposition (SVD). We have borrowed the idea from [41] with significant improvements in implementation and results. The work also goes a step further wherein multiple watermarks inserted and combined with compressive sensing (CS) theory [42-43].

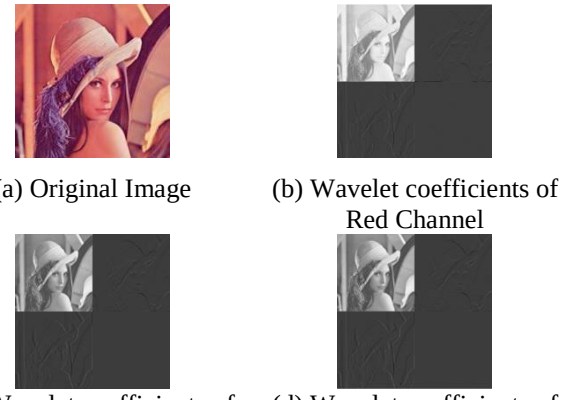
In this proposed algorithm, singular value of approximation wavelet subband of color host image is modified according to the singular value of CS measurements of multiple watermarks. This proposed algorithm offers good security and high payload capacity. In this algorithm, CS theory is applied to DCT coefficients of multiple watermarks information which are inserted into the color host image. In this proposed algorithm, Gaussian measurement matrix A is applied to the DCT coefficients of watermark image to get CS measurements of watermark image. The proposed algorithm is analyzed using various color host images and multiple watermarks for different gain factors. The orthogonal matching pursuit (OMP) [44] algorithm is used for extraction of watermark image from extracted CS measurements at detector side. This algorithm is selected because it has better computational time and easy to implemented.

3 Preliminaries

3.1 Redundant Discrete Wavelet Transform (RDWT)

The most common transform such as discrete wavelet transform is used for watermarking. But DWT has limitation such that downsampling of its subbands [39, 41]. This is cause payload capacity of watermarking algorithms. The DWT is also shift variance which may cause a problem in the extraction of watermark information. So overcome these limitations of DWT in watermarking, researchers are introduced redundant discrete wavelet transform (RDWT) for watermarking. The RDWT provides shift invariance which is better for extraction of watermark information at detector side. The RDWT is eliminated downsampling and the upsampling process of discrete wavelet transform. This transform provides more robust process than DWT. When RDWT is applied on any color image which decomposed the image into various coefficients are shown in Figure 1.

The different between discrete wavelet transform (DWT) and redundant discrete wavelet transform (RDWT) is shown in Figure 2 [39, 41].



(a) Original Image (b) Wavelet coefficients of Red Channel

(c) Wavelet coefficients of Green Channel (d) Wavelet coefficients of Blue Channel

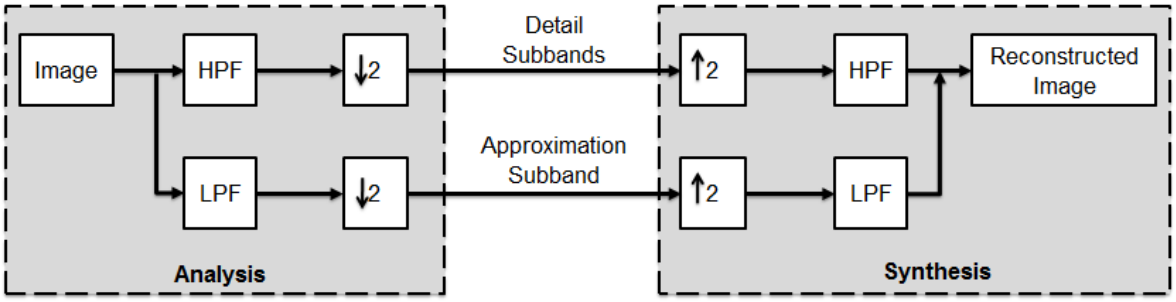
Figure 1: Wavelet Coefficients of RDWT for Color Image.

3.2 Singular Value Decomposition (SVD)

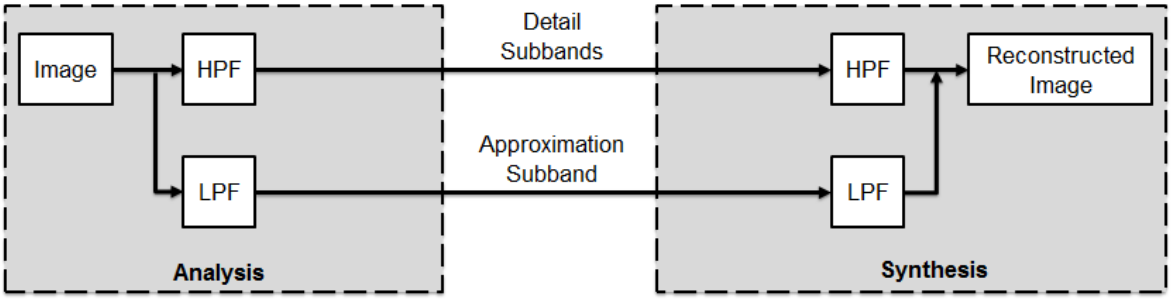
The singular value decomposition is a linear algebra tool which decomposes the image into three different matrices such as singular value matrix, two orthogonal matrices such as U matrix and V matrix. The singular value matrix has non-negative values and diagonally place in the matrix. The singular value has sparsity and stable property which is suitable for watermarking and compression sensing. This value is less effect on human visualization capacity when it is modified. When SVD is applied on any image is shown in Figure 3.

3.3 Compressive Sensing (CS) theory

An image f can become sparse image when only a few non-zero elements are presented in the image. The image f can be converted into a sparse image by applying image



(a)



(b)

Figure 2: (a) DWT Analysis and Synthesis for Image (b) RDWT analysis and Synthesis for Image.

transform basis matrix. The image has x non-zero coefficients (sparse coefficients) are represented as

$$x = \Psi(f) \tag{1}$$

where x is the sparse coefficients, Ψ is the transform basis matrix.

The CS measurements of image using compressive sensing represented by using following equation [42, 43]

$$y = A \times x \tag{2}$$

Where y is the sparse measurements, A is known as measurement matrix.

To reconstruction of an image from CS measurements, various CS recovery algorithms are available in the literature [42-45]. A greedy algorithm such as orthogonal matching pursuit (OMP) is used which is introduced and designed by Tropp et al. [44]. The more detail on OMP algorithm is given in next subsection. It is used in this paper for the extraction of sparse coefficients from CS measurements. It can be mathematically explained using below equation:

$$x^* = OMP(y, A) \tag{3}$$

Where x^* is extracted sparse coefficients which are extracted from the CS measurements y.

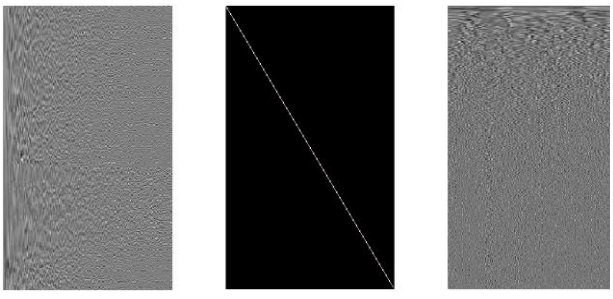


Figure 3: SVD Matrices of Image: U Matrix (left), S Matrix (middle), V Matrix (right).

3.4 Orthogonal Matching Pursuit (OMP) algorithm

The Orthogonal Matching pursuit (OMP) algorithm is introduced and designed by J. Tropp and A. Gilbert in 2007 [44]. This algorithm is a greedy algorithm which is used for extraction of sparse coefficients from the CS measurements. The OMP algorithm is defined by three basic steps such as matching, orthogonal projection, and residual updating. The output of OMP algorithm is one non-zero sparse coefficient in each iteration. The OMP algorithm extracted sparse coefficients x from $y=Ax$. The mathematical steps for OMP algorithm are described in below steps:

- **Input:** CS measurements $\boldsymbol{y}$, Measurement matrix $\boldsymbol{A}$
- **Initialization:** index $\boldsymbol{I = A}$, residual $\boldsymbol{r = y}$, sparse representation $\boldsymbol{\theta = 0 \in Rm}$.
- **Step 1:** Initialize the residual $r_0 = y$ and initialized the set of selected variable $x(c_0) = \phi$. Let iteration counter i = 1.

$$\max_t \left| x_t^{'} r_{i-1} \right| \tag{4}$$

- **Step 2:** Find the variable x_t that solves the maximization problem below using equation (4) and add the variable x_{ti} to the set of selected variables. Update C_i using equation (5).

$$C_i = C_{i-1} \cup \{t_i\} \quad (5)$$

- **Step 3:** Let P_i which is given below equation (6) denote the projection onto the linear space spanned by the elements of $x(c_i)$. Then update residual r using equation (7).

$$P_i = x(C_i)(x(C_i)^{'} x(C_i))^{-1} x(C_i)^{'} \quad (6)$$

$$r_i = (I - P_i)y \quad (7)$$

- **Step 4:** It the stopping condition is achieved, stop the algorithm. Otherwise, set $i = i + 1$ and return to step 2.
- **Output:** Sparse Coefficients ***x*.**

The solution of equation 6 is getting by least square optimization method. The value of projection *Pi* is taken as extracted sparse coefficients *x*. The value of extracted sparse coefficients depends on linear projection between CS measurements vector and measurement matrix-vector. When both vectors have equal value then the output is zero because the projection is zero. So every time, the output of OMP algorithm is non-zero coefficients.

4 Watermarking algorithm using RDWT-SVD and CS theory

Today's world, the data size of multimedia is also increasing day by day. Also, the existed watermarking techniques are embedded multiple watermarks information in host image but the size of watermark information is few bits or less than the size of the host image. Thus, the new watermarking algorithm should have more secure and higher payload capacity. So, in the proposed algorithm CS is used for providing security to multiple watermarks information before embedding. In the proposed algorithm, redundant DWT is applied on RGB channel of color host image to get wavelet coefficients of RGB channel of the color host image. Then SVD is applied on these coefficients to get the singular value of RGB channel of the color host image. Then the singular value of approximation subbands is chosen for CS measurements embedding because of these coefficients are less effect by watermarking attacks. The CS measurements of multiple watermarks information are generated using DCT and Gaussian measurement matrix. In this section, the multiple watermark embedding procedure and extraction procedure of proposed algorithm are described.

4.1 Multiple watermark embedding procedure

The multiple watermark images are transformed into the sparse domain using discrete cosine transform (DCT).

- Take color host image *IH* and compute the size of the host image. Then color image decomposed into R channel, G channel and B channel.

The CS measurements *y* of watermark images is generated using compressive sensing with Gaussian measurement matrix. The singular value of CS measurements *y* of watermark images is embedded into the singular value of LL subband of RGB channel of the color host image. Figure 4 shows the framework for the proposed embedding procedure and the mathematical steps for multiple watermark embedding are given below.

- Take multiple watermarks w_1, w_2, w_3 and compute the size of watermarks. Apply DCT on watermark 1, watermark 2 and watermark 3 to get DCT coefficients of watermark 1, watermark 2 and watermark 3, respectively.

$$D_1 = dct(w_1)$$
$$D_2 = dct(w_2) \quad (8)$$
$$D_3 = dct(w_3)$$

  In above equation, w_1, w_2, *and* w_3 are watermark 1, watermark 2 and watermark 3, respectively; D_1, D_2, *and* D_3 are DCT coefficients of watermark 1, watermark 2 and watermark 3, respectively.
- Then generate CS measurements of watermark 1, watermark 2 and watermark 3 using Compressive sensing procedure. The Gaussian measurement matrix is generated using zero mean and one variance.

$$y_1 = A \times D_1$$
$$y_2 = A \times D_2 \quad (9)$$
$$y_3 = A \times D_3$$

  In above equation, y_1, y_2, *and* y_3 are CS measurements of watermark 1, watermark 2 and watermark 3, respectively; *A* is Gaussian measurement matrix; D_1, D_2, *and* D_3 are DCT coefficients of watermark 1, watermark 2 and watermark 3, respectively.
- Apply SVD on CS measurements of watermark 1, watermark 2 and watermark 3 to get the singular value of CS measurements of watermark 1, watermark 2 and watermark 3, respectively.

$$[U_{Y1}, S_{Y1}, V_{Y1}] = svd(y_1)$$
$$[U_{Y2}, S_{Y2}, V_{Y2}] = svd(y_2) \quad (10)$$
$$[U_{Y3}, S_{Y3}, V_{Y3}] = svd(y_3)$$

  In above equation, S_{Y1}, S_{Y2}, and S_{Y3} are the singular value of CS measurements of watermark 1, watermark 2 and watermark 3, respectively; y_1, y_2, *and* y_3 are CS measurements of watermark 1, watermark 2 and watermark 3, respectively.

- Apply RDWT on R channel, G channel, and B channel of host image to get wavelet coefficients of the host image.

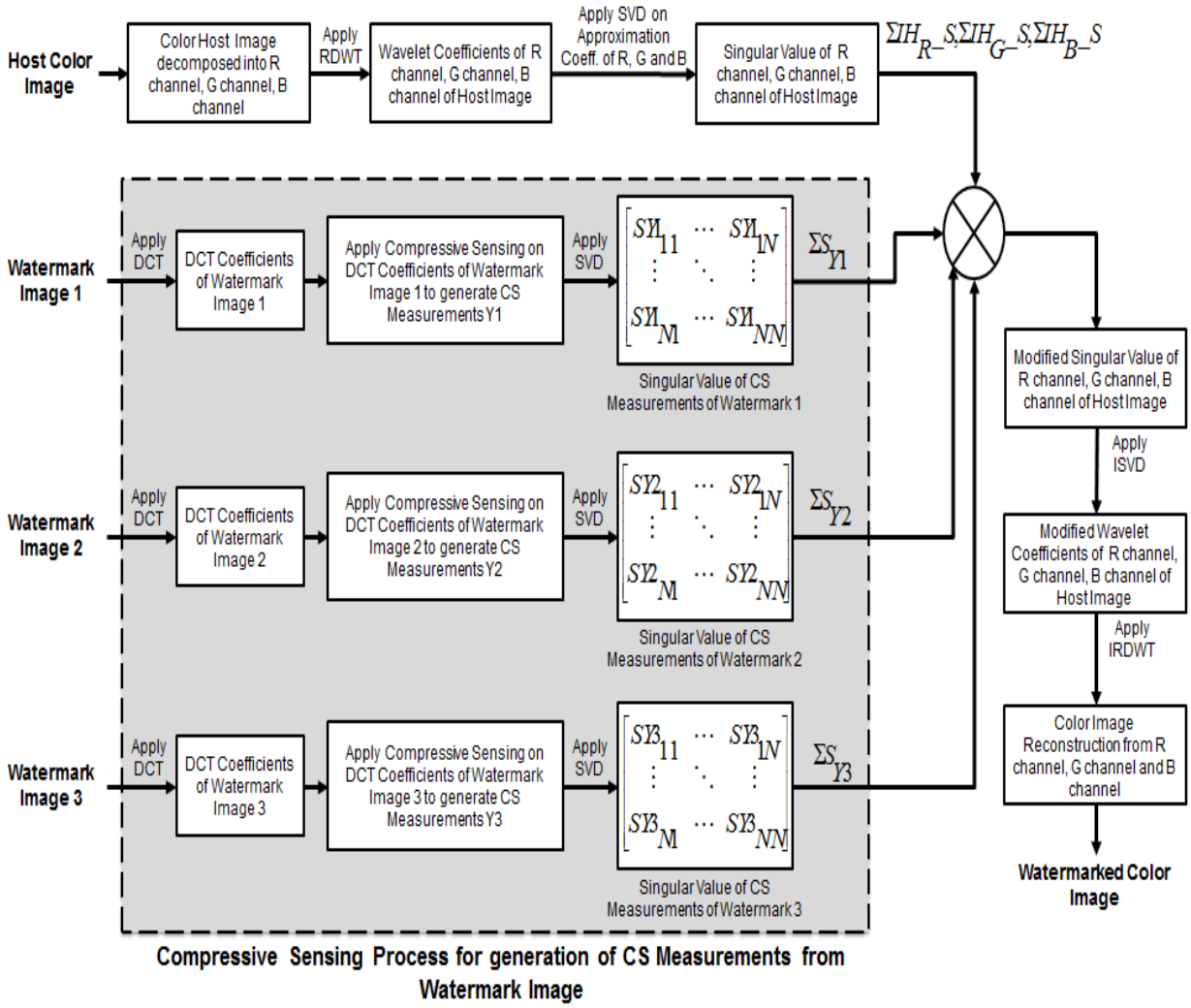


Figure 4: Framework for Proposed Embedding Procedure.

$$[LL1, LH1, HL1, HH1] = RDWT(IH\_\mathrm{Re}d)$$
$$[LL2, LH2, HL2, HH2] = RDWT(IH\_Green) \quad (11)$$
$$[LL3, LH3, HL3, HH3] = RDWT(IH\_Blue)$$

In above equation, *IH_Red*, *IH_Green*, and *IH_Blue* are R channel, G channel and B channel of the color host image, respectively.

- Then SVD is applied on LL subband to get the singular value of approximation wavelet coefficients of the host image.

$$[IH_R\_U, IH_R\_S, IH_R\_V] = svd(LL1)$$
$$[IH_G\_U, IH_G\_S, IH_G\_V] = svd(LL2) \quad (12)$$
$$[IH_B\_U, IH_B\_S, IH_B\_V] = svd(LL3)$$

In above equation, IH_R_S, IH_G_S, and IH_B_S are the singular value of LL subband of R channel, G channel and B channel of the color host image, respectively.

- The singular value of approximation wavelet coefficients of host image is modified according to the singular value of CS measurements of watermark1, watermark 2 and watermark 2 using gain factor.

$$S1 = IH_R\_S + (k \times S_{Y1})$$
$$S2 = IH_G\_S + (k \times S_{Y2}) \quad (13)$$
$$S3 = IH_B\_S + (k \times S_{Y3})$$

In above equation, *S1*, *S2*, and *S3* have modified the singular value of LL subband of R channel, G channel and B channel of the color host image, respectively; *k* is a gain factor.

- Apply inverse SVD on modified singular value to get modified LL subband of R channel, G channel and B channel of the color host image.

$$new\_LL1 = IH_R\_U * S1 * IH_R\_V^{'}$$
$$new\_LL2 = IH_G\_U * S2 * IH_G\_V^{'} \quad (14)$$
$$new\_LL3 = IH_B\_U * S3 * IH_B\_V^{'}$$

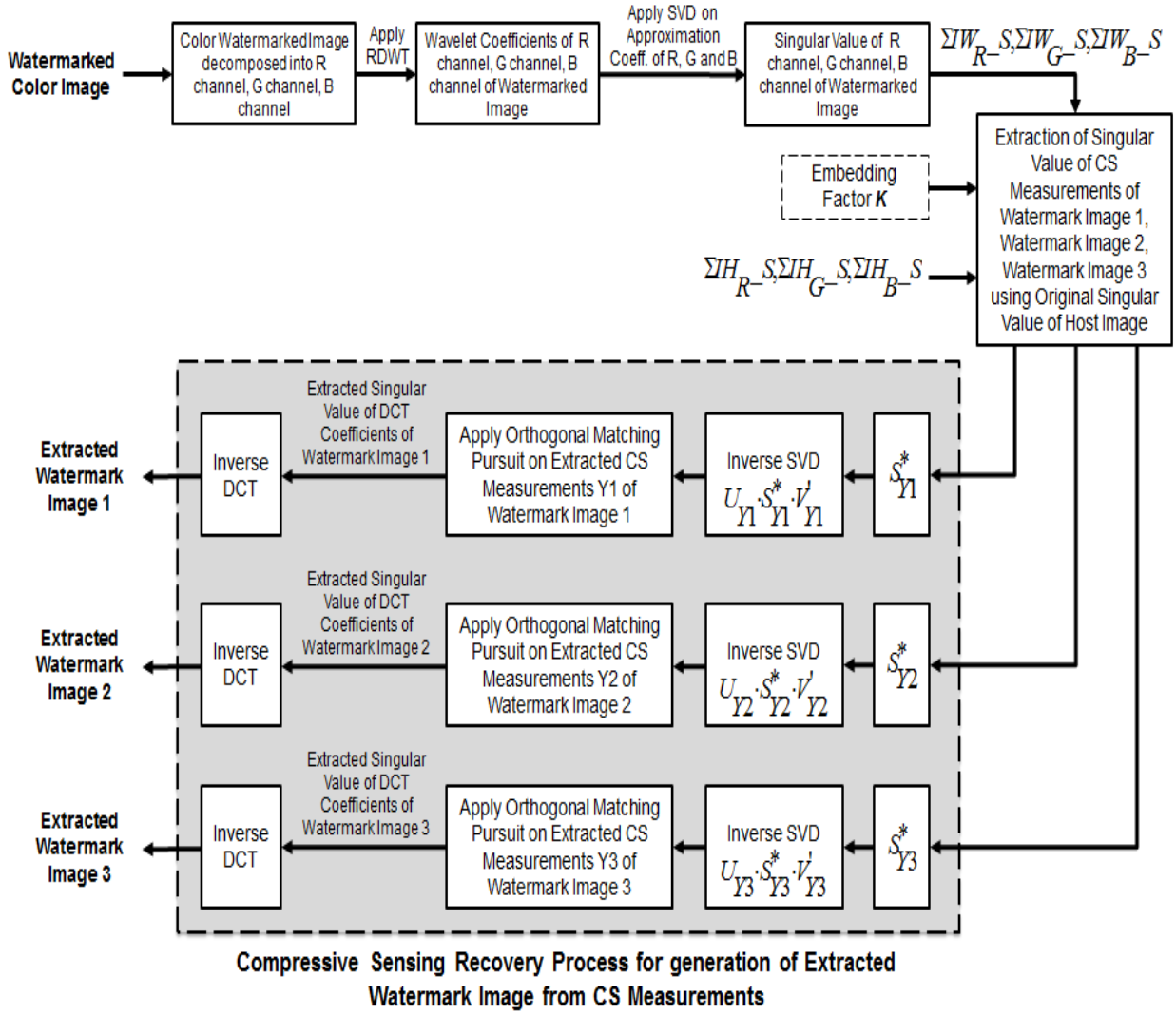


Figure 5: Framework for Proposed Extraction Procedure.

In above equation, *new_LL1*, *new_ LL2*, and *new_LL3* have modified LL subband of R channel, G channel and B channel of the color host image, respectively.

- The inverse RDWT is applied on modified LL subband with unmodified subbands to get modified R channel, G channel and B channel of the color host image.

$$WA\_R = IRDWT(new\_LL1, LH1, HL1, HH1)$$
$$WA\_G = IRDWT(new\_LL2, LH2, HL2, HH2) \quad (15)$$
$$WA\_B = IRDWT(new\_LL3, LH3, HL3, HH3)$$

In above equation, *WA_R*, *WA_G*, and *WA_B* are modified R channel, G channel and B channel of the color host image, respectively.

- Finally, the color image reconstruction is applied on modified RGB channels to get watermarked color image *IW*.

4.2 Multiple watermark extraction procedure

For extraction of the watermark, from watermarked image, the measurement matrix, orthogonal matrices U, V of CS measurements is required. The OMP algorithm [44] is used to the extraction of DCT coefficients of multiple watermarks from its extracted CS measurements values. Figure 5 shows the block diagram for the proposed extraction procedure and the mathematical steps for watermark extraction are given below.

- Take watermarked image and decomposed into R channel, G channel, and B channel. Apply redundant DWT on the watermarked image *IW*, to get the modified wavelet coefficients of R channel, G channel and B channel, respectively.

$$[LLW1, LHW1, HLW1, HHW1] = RDWT(IWR)$$
$$[LLW2, LHW2, HLW2, HHW2] = RDWT(IWG) \quad (16)$$
$$[LLW3, LHW3, HLW3, HHW3] = RDWT(IWB)$$

In above equation, *IWR*, *IWG*, and *IWB* are R channel, G channel and B channel of color watermarked image, respectively.

- Apply SVD on LL subband of R channel, G channel and B channel of watermarked image to get the singular value of R channel, G channel and B channel of watermarked image.

$$[IW_{R\_}U, IW_{R\_}S, IW_{R\_}V] = svd(LLW1)$$
$$[IW_{G\_}U, IW_{G\_}S, IW_{G\_}V] = svd(LLW2) \quad (17)$$
$$[IW_{B\_}U, IW_{B\_}S, IW_{B\_}V] = svd(LLW3)$$

In above equation, IW_R_S, IW_G_S, and IW_B_S are the singular value of LL subband of R channel, G channel and B channel of color watermarked image, respectively.

- Extract singular value of CS measurements of multiple watermarks using singular value of RGB channel of host image and singular value of RGB channel of watermarked image with help of gain factor as

$$S^*_{Y1} = (IW_{R\_}S - IH_{R\_}S)/k$$
$$S^*_{Y2} = (IW_{G\_}S - IH_{G\_}S)/k \quad (18)$$
$$S^*_{Y3} = (IW_{B\_}S - IH_{B\_}S)/k$$

In above equation, S^*_{Y1}, S^*_{Y2}, and S^*_{Y3} are extracted singular value of CS measurements of watermark 1, watermark 2 and watermark 3, respectively.

- Then apply inverse SVD on extracted singular value with original U, V to get extracted CS measurements of multiple watermarks information.

$$Ey_1 = W_{1\_}U * EW_{1\_}S * W_{1\_}V^{'}$$
$$Ey_2 = W_{2\_}U * EW_{2\_}S * W_{2\_}V^{'} \quad (19)$$
$$Ey_3 = W_{3\_}U * EW_{3\_}S * W_{3\_}V^{'}$$

In above equation, Ey_1, Ey_2, *and* Ey_3 are extracted CS measurements of watermark 1, watermark 2 and watermark 3, respectively.

- The OMP algorithm is applied on extracted CS measurements of multiple watermarks information to get DCT coefficients of multiple watermarks.

$$D^*_1 = OMP(Ey_1, A)$$
$$D^*_2 = OMP(Ey_2, A) \quad (20)$$
$$D^*_3 = OMP(Ey_3, A)$$

In above equation, D^*_1, D^*_2, *and* D^*_3 have extracted DCT coefficients of watermark 1, watermark 2 and watermark 3, respectively; A is a measurement matrix.

- Finally applied inverse DCT on extracted DCT coefficients to get multiple watermarks at detector side.

$$w^*_1 = idct(D^*_1)$$
$$w^*_2 = idct(D^*_2) \quad (21)$$
$$w^*_3 = idct(D^*_3)$$

In above equation, w^*_1, w^*_2, *and* w^*_3 are extracted watermark 1, extracted watermark 2 and extracted watermark 3 at detector side, respectively.

5 Results and discussion

The testing of proposed algorithm using various types of images with quality measures are discussed in this section. The various test images and watermarks are discussed in subsection 5.1. The quality measures such as PSNR, NCC, and payload capacity for proposed algorithm is discussed in subsection 5.2. The performance analysis of proposed algorithm for multiple watermarks is discussed in subsection 5.3. The performance analysis of proposed algorithm for multiple biometric watermarks is discussed in subsection 5.4. The comparison of proposed algorithm with existed algorithms is discussed in subsection 5.5.

5.1 Test images and watermarks

(a) (b)

Figure 6: Test Host Color Images (a) Lena (b) Mandril.

The performance of any watermarking scheme varies with different types of images. Therefore, in this paper, two different types of host color images such as Lena image and mandril image are used. In Figure 6, Lean host image and mandril host image have a size of 176×176 pixels and 128×128 pixels, respectively. The various type of watermarks information is taken in this paper. Figure 7 shows various standard watermark images with various frequency coefficients. The cameraman watermark image (which has low frequency coefficients), peppers watermark image (which has middle-frequency coefficients) and Goldhill watermark image (which has high frequency coefficients) have a size of 176×176 pixels. Figure 8 shows various biometric watermarks images. The fingerprint watermark image, iris watermark image, and sign watermark image have a size of 128×128 pixels.

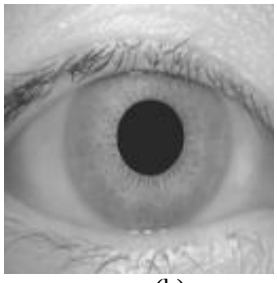
The performance of proposed algorithm is carried out for different gain factor. The analysis of proposed algorithm is carried out for various watermarking attacks such as JPEG compression; noise addition such as Gaussian noise, Salt-Pepper noise, and speckle noise;

(a) (b)

(c)

Figure 7: Test Standard Watermark Images
(a) Cameraman (b) Peppers (c) Goldhill.

(a)



(b)

(c)

Figure 8: Test Biometric Watermark Images
(a) Fingerprint (b) Iris (c) Sign.

filter attacks such as Mean, median, sharpen and Gaussian low pass filter; geometric attacks such as histogram equalization, rotation, and cropping.

5.2 Quality measures

The perceptual quality of watermarked image is measured by Peak Signal to Noise Ratio (PSNR) [46] and the mathematical equation of PSNR is given in below

$$PSNR = 10\log_{10}\frac{255^2}{MSE} \tag{22}$$

In above equation, MSE is defined as mean square error and given by

$$MSE = \frac{1}{M \times N}\sum_{x=1}^{M}\sum_{y=1}^{N}\{I(x,y) - I^*(x,y)\}^2 \tag{23}$$

In above equation, I and I^* is original host image and watermarked image respectively.

The MSE is measured in general scale while PSNR is measured in logarithmic scale. The high value of PSNR is indicated more imperceptibility of watermarking scheme. The normalized cross correlation (NCC) [46] is used to measure the similarity between original watermark image and extracted watermark image. The mathematic equation for NCC is given in below

$$NCC = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N} w(x,y) \times w*(x,y)}{\sum_{x=1}^{M}\sum_{y=1}^{N} w^2(x,y)} \tag{24}$$

In above equation, w is original watermark image and w^* is extracted watermark image.

The NCC value lies in 0 to 1. When NCC value is 1 then it is indicated the extracted watermark image is exactly similar to the original watermark image. But NCC value is 0 then it is indicated that the extracted watermark image is not similar to the original watermark image. In this paper, PSNR is used for measurement of imperceptibility of proposed watermarking algorithm. NCC is used for measurement of robustness and security of proposed watermarking algorithm.

The payload capacity of any watermarking system can be defined as the amount of watermark information is embedded into host medium. The payload capacity can be calculated by a number of bits embedded in host pixels or a ratio of the size of the watermark to the size of host medium. In this paper, payload capacity is calculated using below equation.

$$PC = \frac{SizeofWatermark}{SizeofHost} \tag{25}$$

In above equation, PC is payload capacity of the watermarking algorithm; the size of watermark and host in pixels.

5.3 Performance analysis of proposed algorithm for multiple watermarks

In the proposed algorithm, CS measurements are generated using DCT coefficients of multiple watermark images. In this proposed algorithm, the wavelet coefficients of R channel, G channel and B channel of color host image are generated using db1 or haar wavelet. The db1 or haar wavelet is basic wavelet, simplest, asymmetric and orthogonal as well as bi-orthogonal in nature. These wavelets are commonly used in watermarking. The singular value of CS measurements

(a) Watermarked Lena Host Image with Multiple Standard Watermark Images

(b) Extracted Cameraman Watermark Image

(c) Extracted Peppers Watermark Image

(d) Extracted Goldhill Watermark Image

Figure 9: Results of Proposed Algorithm for Multiple Standard Watermark Images.

is embedded into the singular value of LL subband of the color host image.

Figure 9 shows the watermarked Lena host image with multiple watermarks and extracted standard watermark images without application of watermarking attacks on watermarked color image using gain factor 0.002 and db1 wavelet.

Table 1 shows PSNR and NCC values for multiple standard watermark images for different gain factor without application of watermarking attacks. The performance analysis of proposed algorithm for multiple standard watermark images against various watermarking attacks is carried out for different gain factor. Table 2 shows the performance of proposed algorithm for multiple standard watermark images under JPEG compression and Gaussian noise addition attack.

Table 3 shows the performance of proposed algorithm for multiple standard watermark images under noise addition attacks such as salt-pepper noise and speckle noise. Table 4 shows the performance of proposed algorithm for multiple standard watermark images under various filters such as median and mean.

Table 5 shows the performance of proposed algorithm for multiple standard watermark images under Gaussian low pass filter attack and sharpening attack. Table 6 and 7 shows the performance of proposed algorithm for multiple standard watermark

Table 1: PSNR and NCC values of Proposed Algorithm for Multiple Standard Watermark Images without Application of Watermarking Attacks.

Gain Factor	PSNR (dB)	NCC 1 for Cameraman Watermark Image	NCC 2 for Peppers Watermark Image	NCC 3 for Goldhill Watermark Image
0.002	43.56	0.9703	0.9822	0.9963
0.003	39.94	0.9555	0.9925	0.9963
0.004	37.44	0.9650	0.9897	0.9904
0.005	34.77	0.9636	0.9927	0.9861

Table 2: Performance of Proposed Algorithm for Multiple Standard Watermark Images under JPEG Compression Attack and Gaussian Noise Addition Attack.

Gain Factor	JPEG Compression (Q = 50)			Gaussian Noise (Variance = 0.001)		
	NCC 1	NCC 2	NCC 3	NCC 1	NCC 2	NCC 3
0.002	0.9693	0.9844	0.9884	0.9785	0.9917	0.9932
0.003	0.9627	0.9927	0.9900	0.9595	0.9953	0.9947
0.004	0.9644	0.9889	0.9893	0.9657	0.9924	0.9833
0.005	0.9690	0.9921	0.9929	0.9637	0.9903	0.9904

Table 3: Performance of Proposed Algorithm for Multiple Standard Watermark Images under JPEG Compression Attack and Salt-Pepper Noise and Speckle Noise Addition Attack.

Gain Factor	Salt-Pepper Noise (Variance = 0.005)			Speckle Noise (Variance = 0.004)		
	NCC 1	NCC 2	NCC 3	NCC 1	NCC 2	NCC 3
0.002	0.9585	0.9843	1.0000	0.9674	0.9876	0.9987
0.003	0.9564	0.9907	0.9874	0.9742	0.9888	0.9946
0.004	0.9656	0.9904	0.9884	0.9662	0.9880	0.9905
0.005	0.9676	0.9904	0.9875	0.9694	0.9925	0.9873

Table 4: Performance of Proposed Algorithm for Multiple Standard Watermark Images under Median Filter Attack and Mean Filter Attack.

Gain Factor	Median Filter (Size of Filter Mask = 3×3)			Mean Filter (Size of Filter Mask = 3×3)		
	NCC 1	NCC 2	NCC 3	NCC 1	NCC 2	NCC 3
0.002	0.9584	0.9831	0.9948	0.9864	0.9934	0.9914
0.003	0.9614	0.9930	0.9975	0.9599	0.9939	0.9962
0.004	0.9664	0.9895	0.9918	0.9642	0.9941	0.9966
0.005	0.9636	0.9941	0.9923	0.9665	0.9884	0.9847

Table 5: Performance of Proposed Algorithm for Multiple Standard Watermark Images under Gaussian Low Pass Filter Attack and Sharpening Attack.

Gain Factor	Gaussian Low Pass Filter (Size of Filter Mask = 3×3)			Sharpening Attack		
	NCC 1	NCC 2	NCC 3	NCC 1	NCC 2	NCC 3
0.002	0.9596	0.9798	0.9963	0.9461	0.9804	0.9989
0.003	0.9572	0.9914	0.9874	0.9577	0.9900	0.9898
0.004	0.9658	0.9854	0.9846	0.9693	0.9872	0.9908
0.005	0.9668	0.9917	0.9864	0.9674	0.9938	0.9888

images under rotation attack, cropping attack, and histogram equalization attack.

The NCC value is above 0.94 shown in Table 2 to 7 in the case of different type of watermarking attacks for color host image with multiple standard watermark images. This situation indicated that the algorithm can provide robustness against various type of watermarking attacks for multiple standard watermark images.

5.4 Performance analysis of proposed algorithm for multiple biometric watermarks

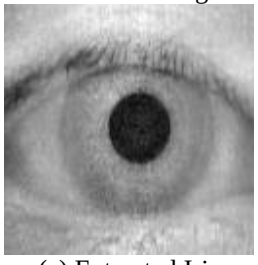
In the proposed algorithm, CS measurements are generated using DCT coefficients of multiple biometric watermark images. In this proposed algorithm, the wavelet coefficients of R channel, G channel and B channel of color host image are generated using db1 wavelet. The singular value of CS measurements is embedded into the singular value of LL subband of the color host image.

Figure 10 shows the watermarked mandril host image with multiple biometric watermarks and extracted biometric watermark images without application of watermarking attacks on watermarked color image using embedding factor 0.002 and db1 wavelet.

Table 8 shows PSNR and NCC values for multiple biometric watermarks for different embedding factor without application of watermarking attacks. The performance analysis of proposed algorithm for multiple biometric watermark images against various watermarking attacks is carried out for different gain factor. Table 9 shows the performance of proposed algorithm for multiple biometric watermark images under JPEG compression and Gaussian noise addition attack.

(a) Watermarked Mandril Host Image with Multiple Biometric Watermark Images

(b) Extracted Fingerprint Watermark Image



(c) Extracted Iris Watermark Image

(d) Extracted Sign Watermark Image

Figure 10: Results of Proposed Algorithm for Multiple Biometric Watermark Images.

Table 10 shows the performance of proposed algorithm for multiple biometric watermark images under noise addition attacks such as salt-pepper noise and speckle noise. Table 11 shows the performance of proposed algorithm for multiple biometric watermark images under various filters such as median and mean.

Table 12 shows the performance of proposed algorithm for multiple biometric watermark images under Gaussian low pass filter attack and sharpening

Table 9: Performance of Proposed Algorithm for Multiple Standard Watermark Images under Rotation Attack and Cropping Attack.

Gain Factor	Rotation Attack			Cropping Attack		
	NCC 1	NCC 2	NCC 3	NCC 1	NCC 2	NCC 3
0.002	0.9864	0.9891	0.9929	0.9633	0.9804	0.9974
0.003	0.9749	0.9884	0.9878	0.9651	0.9920	0.9891
0.004	0.9679	0.9913	0.9958	0.9666	0.9879	0.9908
0.005	0.9637	0.9901	0.9813	0.9657	0.9932	0.9951

Table 10: Performance of Proposed Algorithm for Multiple Standard Watermark Images under Histogram Equalization Attack.

Gain Factor	Histogram Equalization Attack		
	NCC 1	NCC 2	NCC 3
0.002	0.9400	0.9829	0.9989
0.003	0.9604	0.9919	0.9890
0.004	0.9710	0.9950	0.9855
0.005	0.9677	0.9919	0.9905

Table 11: PSNR and NCC values of Proposed Algorithm for Multiple Biometric Watermark Images without Application of Watermarking Attacks.

Gain Factor	PSNR (dB)	NCC 1 for Fingerprint Watermark Image	NCC 2 for Iris Watermark Image	NCC 3 for Sign Watermark Image
0.002	43.15	0.9509	0.9993	0.9714
0.003	39.17	0.9612	1.0000	0.9726
0.004	36.43	0.9653	0.9981	0.9717
0.005	34.19	0.9601	0.9990	0.9700

Table 12: Performance of Proposed Algorithm for Multiple Biometric Watermark Images under JPEG Compression Attack and Gaussian Noise Addition Attack.

Gain Factor	JPEG Compression (Q = 50)			Gaussian Noise (Variance = 0.001)		
	NCC 1	NCC 2	NCC 3	NCC 1	NCC 2	NCC 3
0.002	0.9474	0.9919	0.9753	0.9435	0.9946	0.9742
0.003	0.9603	1.0000	0.9747	0.9595	1.0000	0.9763
0.004	0.9631	0.9992	0.9726	0.9672	0.9988	0.9742
0.005	0.9572	0.9971	0.9689	0.9599	0.9961	0.9696

Table 13: Performance of Proposed Algorithm for Multiple Biometric Watermark Images under JPEG Compression Attack and Salt-Pepper Noise and Speckle Noise Addition Attack.

Gain Factor	Salt-Pepper Noise (Variance = 0.005)			Speckle Noise (Variance = 0.004)		
	NCC 1	NCC 2	NCC 3	NCC 1	NCC 2	NCC 3
0.002	0.9436	0.9985	0.9725	0.9498	0.9959	0.9772
0.003	0.9598	0.9897	0.9725	0.9603	1.0000	0.9745
0.004	0.9622	0.9983	0.9712	0.9628	0.9956	0.9713
0.005	0.9583	0.9994	0.9677	0.9589	0.9951	0.9707

Table 14: Performance of Proposed Algorithm for Multiple Biometric Watermark Images under Median Filter Attack and Mean Filter Attack.

Gain Factor	Median Filter (Size of Filter Mask = 3×3)			Mean Filter (Size of Filter Mask = 3×3)		
	NCC 1	NCC 2	NCC 3	NCC 1	NCC 2	NCC 3
0.002	0.9460	1.0000	0.9731	0.9447	1.0000	0.9718
0.003	0.9572	1.0000	0.9741	0.9592	0.9908	0.9738
0.004	0.9640	0.9971	0.9724	0.9634	0.9962	0.9728
0.005	0.9579	0.9942	0.9658	0.9586	1.0000	0.9695

Table 15: Performance of Proposed Algorithm for Multiple Biometric Watermark Images under Gaussian Low Pass Filter Attack and Sharpening Attack.

Gain Factor	Gaussian Low Pass Filter (Size of Filter Mask = 3×3)			Sharpening Attack		
	NCC 1	NCC 2	NCC 3	NCC 1	NCC 2	NCC 3
0.002	0.9443	1.0000	0.9741	0.9534	1.0000	0.9755
0.003	0.9591	1.0000	0.9746	0.9565	1.0000	0.9735
0.004	0.9604	1.0000	0.9705	0.9621	0.9973	0.9706
0.005	0.9556	0.9945	0.9712	0.9566	0.9996	0.9708

Table 16: Performance of Proposed Algorithm for Multiple Biometric Watermark Images under Rotation Attack and Cropping Attack.

Gain Factor	Rotation Attack			Cropping Attack		
	NCC 1	NCC 2	NCC 3	NCC 1	NCC 2	NCC 3
0.002	0.9512	0.9949	0.9749	0.9455	1.0000	0.9717
0.003	0.9630	0.9955	0.9719	0.9640	1.0000	0.9736
0.004	0.9646	0.9998	0.9713	0.9615	0.9950	0.9715
0.005	0.9577	0.9995	0.9678	0.9545	0.9998	0.9669

attack. Table 13 and 14 shows the performance of proposed algorithm for multiple biometric watermark images under rotation attack, cropping attack, and histogram equalization attack.

The NCC value is above 0.94 shown in Table 9 to 14 in the case of different type of watermarking attacks for color host image with multiple biometric watermark images. This situation indicated that the algorithm can provide robustness against various type of watermarking attacks for multiple biometric watermark images.

Table 17: Performance of Proposed Algorithm for Multiple Biometric Watermark Images under Rotation Attack and Cropping Attack.

Gain Factor	Histogram Equalization Attack		
	NCC 1	NCC 2	NCC 3
0.002	0.9423	1.0000	0.9730
0.003	0.9601	1.0000	0.9721
0.004	0.9616	0.9940	0.9725
0.005	0.9570	1.0000	0.9693

Table 18: Comparison of Proposed Algorithm with Existed Algorithms.

Techniques	Existed Algorithm in [25]	Existed Algorithm in [41]	Proposed Algorithm
Type of Multiple Watermarking	Successive	-	Composite
Host Medium	Grayscale Image	Grayscale Image	Color Image
Type of Watermark Information	Face, Speech, and Sign	Grayscale Image	Grayscale Image, Fingerprint, Iris and Sign
Number of Watermarks	Three	Single	Three
Used Signal Processing Transform	DWT and DCT	RDWT and SVD	RDWT, SVD, and DCT
CS theory is applied on Watermark Image	No	No	Yes
PSNR (dB)	33.2819	37.52	43.56

After obtaining results of proposed algorithm for multiple watermarks, it is clearing seen that this algorithm can be applied any type of watermark information. The NCC value obtained for various watermarking attacks are above 0.90 which is indicated this algorithm provides robustness against any manipulation. This situation indicated that this algorithm can be used for security of multimedia data when it is transferred over a non-secure channel. This algorithm is also used in a multimodal biometric system where multiple biometric data can be secure when it is transferred over a communication channel between two modules.

5.5 Comparison of proposed algorithm with existed algorithms

The comparison of proposed algorithm with existed algorithms with various features is given in Table 15.

The comparison shows that the CS theory is applied on multiple watermarks in proposed algorithm before embedding which is not applied in existed algorithms. The PSNR value of proposed algorithm is better than existed algorithms in the literature. The proposed algorithm can be used for any multiple watermarks such as standard images and biometric images while algorithm in [25] is used for biometric images and algorithm in [41] is used for the standard image.

6 Conclusion

Nowadays, the payload capacity and multiple watermarks inserted ability of watermarking algorithms are considered as important parameters. So in this paper, multiple watermarks based watermarking algorithm with high payload capacity is presented. A watermarking algorithm using redundant DWT, SVD and compressive sensing for multiple watermarks protection is designed and analyzed for the security of multiple watermarks. The proposed algorithm is also shown application on CS theory for generation of multiple CS measurements of watermark image. The CS theory is providing protection to multiple watermark images before embedding in proposed algorithm. Our algorithm is flexible and may be used for any type of watermark information such as standard image and biometric image. This proposed algorithm can be used for security of multiple watermark information when it is transferred over the non-secure communication channel.

An experimental is implemented using various color host images and watermarks information. Two types of host images and six types of watermark images are used in the experiments. The proposed algorithm has generally overcome the limitation of existed watermarking algorithms. The experiments also show that the proposed algorithm can provide robustness against various watermarking attacks such as JPEG compression, the addition of noise, filter, sharpening, histogram equalization, geometric attack such as rotation, cropping. The proposed algorithm is performed better than existed algorithms available in the literature based on PSNR values.

The limitation of proposed algorithm is that original host image is required at detector side for extraction of watermark information. So this algorithm is a non-blind watermarking algorithm in nature. In the future, the algorithm is designed and analyzed for various data such as speech signal, ECG signal, digital video and digital audio signal. Also, a hardware implementation of proposed algorithm is designed using various DSP platform and FPGA kits.

7 References

[1] R. Thanki and A. Kothari, "Digital Watermarking – Technical Art of Hiding a Message", Intelligent Analysis of Multimedia Information July 2016, pp. 426-460.

[2] F. Thakkar and V. Srivastava, "A Fast Watermarking Algorithm with Enhanced Security using Compressive Sensing and Principle Components and its Performance Analysis against a set of Standard Attacks", Multimedia Tools and Applications, 75(21), November 2016.

[3] A. Gupta and M. Raval, "A Robust and Secure Watermarking Scheme based on Singular Value Replacement", Sadhana, 37(4), August 2012, pp. 425-440.

[4] M. Kamlakar, C. Gosavi and A. Patankar, "Single Channel Watermarking for Video using Block based SVD", International Journal of Advances in Computing and Information Researches, 1(2), April 2012.

[5] M. Ramalingam, "Stego Machine – Video Steganography using Modified LSB Algorithm", World Academy of Science, Engineering and Technology, 74, 2011, pp. 502-505.

[6] R. Paul, "Review of Robust Video Watermarking Techniques", IJCA Special Issue on Computational Science – New Dimensions and Perspectives, NCCSE, 3, 2011, pp. 90-95.

[7] V. Santhi and A. Thangavelu, "DWT SVD Combined Full band Robust Watermarking Technique for Color Images in YUV Color Space", International Journal of Computer Theory and Engineering, 1(4), October 2009.

[8] S. Mostafa, A. Tolba, F. Abdelkader and H. Elhindy, "Video Watermarking Scheme based on Principal Component Analysis and Wavelet Transform", International Journal of Computer Science and Network Security, 9(8), August 2009, pp. 45-52.

[9] A. Essaouabi and E. Ibnelhaj, "A 3D Wavelet based Method for Digital Video Watermarking", Proceedings of the 4th IEEE Intelligent Information Hiding and Multimedia Signal Processing, July 2009.

[10] A. Mansouri, A. Mahmoudi, Aznaveh and F. Azar, "SVD based Digital Image Watermarking using Complex Wavelet Transform", Sadhana, 34(30), June 2009, pp. 393-406.

[11] R. Preda and D. Vizireanu, "Blind Watermarking Capacity Analysis of MPEG2 Coded Video", Proceedings of Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services, Serbia September 2007, pp. 465-468.

[12] R. Dili and E. Mwangi, "An Image Watermarking Method Based on the Singular Value Transformation and the Wavelet Transformation", Proceedings of IEEE AFRICON, 2007, pp. 1-5.

[13] L. Fan and F. Yanmei, "A DWT based Video Watermarking Algorithm Applying DS-CAMA", IEEE Region 10 Conference TENCON 2006, November 2006.

[14] M. El-Gayyar, "Watermarking Techniques – Spatial Domain Digital Rights Seminar", Media Informatics, University of Bonn, Germany, May 2006.

[15] C. Chan and L. Cheng, "Hiding Data in Images by Simple LSB Substitution", Pattern Recognition, 37, 2004, pp. 469-474.

[16] F. Huang and Z. Guan, "A Hybrid SVD-DCT Watermarking Method Based on LPSNR", Pattern Recognition Letters 25, 2004, pp. 1769-1775.

[17] E. Ganic and A. Eskicioglu, "Secure DWT-SVD Domain Image Watermarking Embedding Data in All Frequencies", ACM Multimedia and Security Workshop 2004, 2004, pp. 1-9.

[18] C. Podilchuk and E. Delp, "Digital Watermarking: Algorithms and Applications", IEEE Signal Processing Magazine, 18(4), 2001, pp. 33-46.

[19] M. Ejima and A. Miyazaki, "A Wavelet Based Watermarking for Digital Images and Videos", *IEEE International Conference on Image Processing*, August 2000, pp. 678-681.

[20] J. Hernandez, M. Amado and F. Perez-Gonzalez, "DCT domain Watermarking Techniques for Still Image: Detector Performance Analysis and a New Structure", IEEE Transactions on Image Processing, 9, January 2000, pp. 55-68.

[21] G. Langelaar, I. Setyawan and R. Lagendijk, "Watermarking of Digital Image and Video Data – A State of Art Review", IEEE Signal Processing Magazine, September 2000, pp. 20-46.

[22] I. Cox, J. Kilian, T. Shamoon and F. Leighton, "Secure Spread Spectrum Watermarking for Multimedia", IEEE Transactions on Image Processing, 6(12), December 1997, pp. 1673-1687.

[23] P. Rege, "Biometric Watermarking", National Seminar on Computer Vision and Image Processing, September 2012.

[24] E. Lin and E. Delp, "A Review of Fragile Image Watermarks", In Proceedings of ACM Multimedia and Security Workshop, 1, October 1999, pp. 25-29.

[25] V. Inamdar and P. Rege, "Dual Watermarking Technique with Multiple Biometric Watermarks", Sadhana © Indian Academy of Science, 29(1), February 2014, pp. 3-26.

[26] N. Gavini and S. Borra, "Lossless Watermarking Technique for Copyright Protection of High-Resolution Images", In Region 10 Symposium, 2014 IEEE, April 2014, pp. 73-78.

[27] S. Borra and H. Lakshmi, "Visual Cryptography Based Lossless Watermarking for Sensitive Images", In International Conference on Swarm, Evolutionary, and Memetic Computing, Springer International Publishing, December 2015, pp. 29-39.

[28] S. Borra and D. Swamy, "Visual Secret Sharing Based Digital Image Watermarking", International Journal of Computer Science Issues, 9(3), May 2012, pp. 312-317.

[29] N. Dey, M. Pal and A. Das, "A Session Based Blind Watermarking Technique within the NROI of Retinal Fundus Images for Authentication Using DWT, Spread Spectrum and Harris Corner Detection", International Journal of Modern Engineering Research (IJMER), 2(3), June 2012, pp. 749-757.

[30] N. Dey, S. Samanta, X. Yang, A. Das and S. Chaudhuri, "Optimisation of Scaling Factors in Electrocardiogram Signal Watermarking Using Cuckoo Search", International Journal of Bio-Inspired Computation, 5(5), October 2013, pp. 315-326.

[31] N. Dey, B. Nandi, P. Das, A. Das and S. Chaudhuri, "Retention of Electrocardiogram Features Insignificantly Devalorized as an Effect of Watermarking for a Multimodal Biometric Authentication System", Advances in Biometrics for Secure Human Authentication and Recognition, CRC Press 2013, pp. 175-212.

[32] N. Dey, M. Dey, S. Mahata, A. Das and S. Chaudhuri, "Tamper Detection of Electrocardiographic Signal using Watermarked Bio-hash code in Wireless Cardiology", International Journal of Signal and Imaging Systems Engineering, 8(1/2), 2015, pp. 46-58.

[33] Y. Amar, I. Trabelsi, N. Dey and M. Bouhlel, "Euclidean Distance Distortion Based Robust and Blind Mesh Watermarking", International Journal of Interactive Multimedia and Artificial Intelligence, 4(2), December 2016, pp. 46-51.

[34] B. Surekha, G. Swamy and K. Rao, "A Multiple Watermarking Technique for Images based on Visual Cryptography", International Journal of Computer Applications, 1(11), 2010, pp. 77-81.

[35] N. Dey, G. Dey, S. Chakraborty and S. Chaudhuri, "Feature Analysis of Blind Watermarking Electromyogram Signal in Wireless Telemonitoring", Concepts and Tends in Healthcare Information Systems, Springer International Publishing, 2014, pp. 205-229.

[36] M. Raval and P. Rege, "Discrete Wavelet Transform Based Multiple Watermarking Scheme", Proceedings of the Convergent Technologies for the Asia-Pacific Region, 3, 2003, pp. 935-938.

[37] M. Vatsa, R. Singh and A. Noore, "Feature based RDWT Watermarking for the Multimodal Biometric System", Image and Vision Computing, 27(3), 2009, pp. 293-304.

[38] R. Thanki, R. Kher and D. Vyas, "Analysis of Multiple Users Watermarking in Spatial Domain", International Journals of Computer Science and Telecommunications (IJCST), 2(5), August 2011, pp. 19-22.

[39] T. Hiena, Z. Nakaoa and Y. Chen, "Robust Multi-logo Watermarking by RDWT and ICA", Signal Processing, 86, 2006, pp. 2981–2993.

[40] R. Thanki and K. Borisagar, "Compressive Sensing Based Multiple Watermarking Technique for Biometric Template Protection", International Journal of Image, Graphics and Signal Processing, 7(1), December 2014, 53-60.

[41] S. Lagzian, M. Soryani and M. Fathy, "A New Robust Watermarking Scheme based on RDWT-SVD", International Journal of Intelligent Information Processing, 2(1), March 2011, pp. 22-29.

[42] D. Donoho, "Compressed Sensing", IEEE Transaction on Information Theory, 52(4), April 2006, pp. 1289-1306.

[43] E. Candes, "Compressive Sampling", In Proceedings of the International Congress of Mathematicians, Madrid, Spain, 3, August 2006, pp. 1433-1452.

[44] J. Tropp and A. Gilbert, "Signal Recovery from Random Measurements via Orthogonal Matching Pursuit", IEEE Transactions on Information Theory, 53(12), December 2007, pp. 4655-4666.

[45] M. Duarte and Y. Eldar, "Structured compressed sensing: From theory to applications", IEEE Transactions on Signal Processing, 59(9), 2011, pp. 4053-4085.

[46] M. Kutter and F. Petitcolas, "A fair benchmark for image watermarking systems", Electronic Imaging '99, Security and Watermarking of Multimedia Contents, 3657, January 1999, pp. 1-14.

A Hybrid Approach from Ant Colony Optimization and K-nearest Neighbor for Classifying Datasets Using Selected Features

Enas M. F. El Houby, Nisreen I. R. Yassin and Shaimaa Omran
Systems & Information Department, Engineering Division, National Research Centre, Dokki, Cairo 12311, Egypt
E-mail: enas_mfahmy@yahoo.com, eng_nesrin@hotmail.com, shmomran@gmail.com

Keywords: ant colony optimization, K-nearest neighbor, features selection, heuristic, pheromone.

Received: December 12, 2016

This paper presents an Ant Colony Optimization (ACO) approach for feature selection. The challenge in the feature selection problem is the large search space that exists due to either redundant or irrelevant features which affects the classifier performance negatively. The proposed approach aims to minimize the subset of features used in classification and maximize the classification accuracy. The proposed approach uses several groups of ants; each group selects the candidate features using different criteria. The used ACO approach introduces the datasets to a fitness function that is composed of heuristic value component and pheromone value component. The heuristic information is represented with the Class-Separability (CS) value of the candidate feature. The pheromone value calculation is based on the classification accuracy resulted by adding the candidate feature. K-Nearest Neighbor is used as a classifier. The sequential forward feature selection has been applied, so it selects from the highest recommended features sequentially until the accuracy is enhanced. The proposed approach is applied on different medical datasets yielding promising results and findings. The classification accuracy is increased with the selected features for different datasets. The selected features that achieved the best accuracy for different datasets are given.

Povzetek: Opisan je hibridni pristop optimiranja s pomočjo optimizacije s kolonijami mravelj in metodo k-najbližjih sosedov.

1 Introduction

Real life data processing means having a huge amount of features that need to be analyzed, mined, classified and modeled. Classification is an important process which aims to predict the classes of future data objects. It is an automated process that requires previous knowledge of the datasets to construct a class for each group of relevant features. The aim of building classifier is to find a set of features that gives the best classification accuracy. The classification accuracy is affected by the relevancy of one feature to the other [1]. Redundant and irrelevant features worsen the performance of a classifier. This can be avoided by selecting and grouping relevant features only, thus feature selection reduces the training time and minimizes the feature set and enhances the performance of the classifier [2, 3].

The challenge in feature selection algorithm is to select minimum subset of features by eliminating features that are redundant and irrelevant which may lead the classification process to undesirable results and also removing features that do not provide predictive information. This selection is to be done with no loss of classification accuracy while reducing computation time and cost. Feature selection is an important data preprocessing phase for mining and classifying data. It is a process of selecting the optimum set of features based on a certain specified criterion to construct solid learning models [4-6]. Feature selection algorithms are divided into two categories; the first one covers the filter approach, it is an individual feature ranking approach which ranks features based on statistical methods. The second category covers the wrapper approach, which uses classifiers having classification functions to select those features with high prediction performance [7, 8].

As computation of huge number of features is not feasible; heuristic search methods are needed for feature selection. Many meta-heuristics approaches have been proposed for feature selection, such as nature inspired algorithms which have been used to select features. These algorithms like ant colony optimization have been applied to feature selection as no solid heuristic exist to find optimal feature subset, so it is expected that the ants discover good feature combinations as they proceed through the search space. Such optimization techniques were used for modeling the feature selection as an optimization problem [1, 9].

Based on this idea, in this research an Ant Colony Optimization (ACO) approach for feature selection is applied using different novel search criteria where each group of ants uses a different search criterion such as the standard deviation, the nearest, the furthest, …etc to discover different good feature combinations. In this research work, the nearest and the furthest criteria specifically were implemented. The proposed ACO approach aims to find the minimum set of features that

provide the best classification accuracy. Thus, the objective is to minimize the number of features and the classification error. The next sections are organized as follow: an overview for the previous work related to subject is presented in section 2. An introduction of ACO in feature selection problems is discussed in section 3. The proposed model is described in section 4. The experimental results are presented in section 5 before drawing conclusions and future work in section 6.

2 Related work

The feature selection and classification have become active research areas, as recently several researchers investigated various techniques to address the feature selection and classification problem. Different swarm intelligent optimizations have been used for feature selection and classification in many literatures.

Fong et al. [10] presented a study for feature selection of high dimensional biomedical datasets. They used three meta-heuristic techniques which are the particle swarm optimization (PSO), the wolf search algorithm and the bat algorithm integrated with three classification algorithms to advance the feature selection techniques and thus lowers the classification error rate. The proposed search techniques were applied on 2 biomedical datasets which are Arrhythmia and Micro Mass. Chen, et al. [11] proposed a regression based particle swarm optimization to address the feature selection problem. Regression model was used to find if the feature was selected or not using fitness values for features. Nine data sets from UCI machine learning databases were used for evaluation of the proposed algorithm. Khuat et al. [12] used directed artificial bee colony algorithm to optimize the parameters of the software effort estimation models. The accuracy of the models after optimizing parameters was improved relative to the original models accuracy. Xue et al. [13] introduced two multi-objective algorithms for feature selection based on PSO. The first one was concerned with sorting for PSO and the second one applied the crowding and mutation for PSO. They were implemented on benchmark data sets. Khazaee et al. [14] presented the PSO technique to optimize the input feature subset selection and to set the parameters for a SVM based classifier. The proposed technique was applied on three datasets for three types of electrocardiogram beats. Yeh et al. [15] presented a rule-based classifier that is constructed using simple swarm optimization, to perform the feature selection study on a thyroid gland dataset from UCI databases.

Sivagaminathan et al. [16] introduced a model that used a hybrid method of ant colony optimization and artificial neural networks (ANNs) to select features subset from medical datasets. The heuristic was calculated as a function of cost, so the feature with the lower cost was considered better and selected. Jona et al. [17] proposed a hybrid meta-heuristic search for feature selection in digital mammogram. The proposed search used a hybrid of Ant Colony Optimization (ACO) and Cuckoo Search (CS) which was used to speed local search of ACO. Support Vector Machine (SVM) was used with the ACO to classify the mammogram as normal or abnormal. Asad et al. [18] used ant colony system for features selection of retinal images dataset, a comparative study was conducted among six different features selection heuristics. They concluded that relief heuristic selection is better than the subsets selected by other heuristics. Tallon-Ballesteros et al. [19] proposed the use of Ant System (AS) search technique with two feature subset selection methods which are Correlation-based Feature Selection (CFS) and Consistency-based Feature Selection (CNS). They found that information gain is appropriate heuristic with both CFS and CNS. Dadaneh et al. [20] developed unsupervised probabilistic feature selection using ant colony optimization (UPFS). They decreased redundancy using inter-feature information which shows the similarity between the features. A matrix was used to save pheromone value between each pair of features; it was used to rank features. SVM, K-nearest neighbor (KNN), and naive Bayes (NB) were used as classifiers. Wang et al. [21] proposed a system that adjusts the parameter of ACO using different strategies to understand the parameters of ACO. The parameters included number of ants, pheromone evaporation rate, and exploration probability factor. ACO had been modified by combining it with fuzzy to be used as adaptive method for parameters. Ten UCI and StatLog benchmark data sets had been used to evaluate the performance of the proposed system. Liu et al. [22] used Bee Colony Optimization (BCO), ACO, and PSO to discover the quality of data using approaches to detect attribute outliers in datasets. The same fitness function had been used for the different search strategies. Chen et al. [23] presented an algorithm for feature selection based on ACO which traverse arcs on a directed graph; the heuristic depends on the performance of classifier and the number of selected features. SVM is used as classifier, other classifiers are used for the purpose of comparison, but SVM outperforms the other classifiers.

The introduced literatures in this section cover some of the recent researches concerned with feature selection and classification using swarm inspired algorithms. However the literatures [16-23] used specifically ACO for feature selection. Different classifiers such as ANN, SVM, KNN, and NB had been used; different heuristic such as function of cost, cuckoo search, information gain, performance of classifier and the number of selected features had been applied; different pheromone value calculation methods had been proposed; ACO parameters adjusting had been studied. And other updates had been provided to enhance the performance of ACO in features selection so as to reach the best possible accuracy with the least number of features.

The proposed work added up to these previous findings to enhance the performance of ACO in features selection. Our proposed model provides a novel idea of using different groups of ants which are synchronizing search for different solutions each using a different search criterion to reach the best possible solution for each group. Then the global best solution of all is obtained from the different applied criteria. The selection of features for different groups is done using the same fitness function but with different criteria. The fitness function is

depending on heuristic information term which is represented by Class-separability (CS) and pheromone value term which is updated using a function in the classification accuracy. The features selection is done with considering sequential forward feature selection that the feature with the best fitted fitness value is selected as it improves the classification accuracy of the selected subset; otherwise the feature with the next value is selected. A comparison between the performance of the proposed research and the previous closest work that use the same dataset will be provided in the "Experimental Results and Discussion" section.

3 Ant colony optimization

Artificial swarm intelligence is a population based meta-heuristic technique that is inspired from the behavior of living beings that live in groups, colonies, and flocks. These living organisms interact with each other and with their surrounding environment in a way that can be observed as a behavior pattern. This behavior pattern can be modeled using computers to solve different combinatorial problems. The ACO algorithm is a swarm intelligence technique that mimics real ants' attitude and behavior when searching for and grabbing their food. Ants are blind creatures that have the ability to find the shortest path from their nest to the food source. When searching for the food, ants initially explore the area surrounding their nest in a random manner. After finding the food; the ants carry part of the food and return to their nests. They release a chemical material called pheromone on the ground when moving from the food source location back to their nest. The amount of the pheromone released which mostly depends on the quantity and quality of food guides the other ants to the location of the food, and enables them to find the shortest path from their nest to the food. The ants that use a shorter path first time to grab the food returns to the nest faster releasing larger quantity of pheromone for shorter routes rather than longer routes. Afterwards, the specified shorter path will be preferred almost by all ants and as a result the pheromone starts to evaporate. The probabilistic route selection helps the ants to find the shortest routes and provide flexibility for solving optimization problems [16, 24, 25].

The ACO technique was introduced in the nineties by Dorigo et.al to solve optimization problems as the travelling salesman problem (TSP) [26]. The solution begins from a start node (feature), afterwards the ant moves iteratively from one node (feature) to the other. A probabilistic transition rule is the most widely used function in ACO which is based on the value of the heuristic function η and the pheromone value τ. It is used to iteratively construct a solution. So, the ant moves to an unvisited node (feature) with a probability of:

$$P_i^k(t) = \frac{(\tau_i(t))^{\alpha}(\eta_i(t))^{\beta}}{\Sigma_{j\epsilon N_j^k}(\tau_j(t))^{\alpha}(\eta_j(t))^{\beta}}, \quad j \in N_j^k \qquad (1)$$

Where:

N_j^k is the feasible neighborhood of the ant k, which are the features that ant k has not yet selected and can be chosen. It acts as the memory of the ants.

$\eta_i\ (t)$ is the heuristic information of the feature (i) at the time t.

$\tau_i(t)$ is the pheromone value on the feature (i) at the time t.

α and β are weights that represent the relative impact of the pheromone τ and the heuristic information η, respectively. α and β are parameters that may take real positive values according to the recommendations on parameter setting in [27].

All ants update pheromone level $\tau_i(t)$ with an increase of pheromone quantities, depending on the equations for pheromone updating, which specify how to modify the pheromone value. These equations are determined by:

$$\tau_i(t+1) = \tau_i(t) \times (1-\rho) + \Delta\tau_i(t) \qquad (2)$$

$$\Delta\tau_i^k(t) = \begin{pmatrix} Q & \text{if the feature (i) is chosen by the ant k} \\ 0 & \text{otherwise} \end{pmatrix} \qquad (3)$$

where: ρ is the pheromone evaporation rate of the features ($0 < \rho < 1$), and

$\Delta\tau_i^k$ is the pheromone deposited by the ant k that found the best solution for the current iteration.

In the ACO algorithm, ants exchange information with each other through the pheromone value. Each ant uses information obtained from other ants to provide better solutions. This improves the efficiency of solutions obtained in consecutive iterations. Thus the algorithm can be considered as a multi agent technique [28, 29].

4 The proposed ACO model

The proposed model presents an ACO approach for feature selection. Two objectives are considered; minimizing the number of features used in classification and increasing the classification accuracy. The proposed algorithm is presented in Pseudo code1.

Firstly, the algorithm starts by initializing the following parameters: The number of generations which represents the number of iterations, the number of groups of ants which represent the number of different criteria for features selection, the number of ants which is equivalent to the number of solutions for each criteria (group), maximum number of features that represent maximum allowed number of features that can be selected by each ant to achieve the best possible classification accuracy. (τ_i) which is the pheromone concentration value associated with feature (f_i) and (η_i) is the heuristic value for feature (f_i); (τ) & (η) together form the fitness function terms as shown in eq. (1). α, β are user selected parameters; they represent the relative importance of pheromone and heuristic values respectively. ρ is a user defined parameter that represents the pheromone evaporation or decay rate, it takes a value from 0 to 1. Z is the local pheromone update parameter, it is defined by the user and it takes a value less than ρ.

A set of generations starts after the initialization phase. With each new generation, n groups of ants are formed where G_1, G_2,….G_n are n different groups each having na ants. The first feature selection for each ant is

performed randomly taking into consideration avoiding redundancy between ants of the same group to obtain different possible solution sets. For each selected feature by different ants an initial value for the classification accuracy is obtained for each ant using the KNN algorithm. Using a set of equally initial pheromone values, and the local pheromone update parameter Z, the pheromone of the selected feature is locally updated using eq. (4).

$$\tau_i(t+1) = (1-Z) \times \tau_i(t) + Z \quad (4)$$

For each group of ants, the selection of the subsequent features is done using the same fitness function with different criteria. One used criterion is to select the nearest feature to the previously selected one according to fitness value. Another used criterion is to get the furthest feature to the previously selected one according to fitness value. The fitness function is calculated using eq. (1) having a term representing the pheromone τ and a heuristic term η. The heuristic information is represented with the Class-Separability (CS) value of the feature. All features have equally initial pheromone values and the pheromone of the selected feature is locally updated with eq. (4). By the end of each generation, the pheromone values of the features subsets that are part of the best solution for different groups are globally updated using eq. (5). It is a function in the classification accuracy achieved by selected features subset, so as to increase the features selection opportunity of these features in the future generations.

$$\tau_i(t+1) = (1-\rho) \times \tau_i(t) + \rho * acc \quad (5)$$

Where: ρ is the pheromone evaporation rate of the features ($0 < \rho < 1$), and acc is the classification accuracy.

As mentioned above, the selection of the subsequent feature is done using different criteria. This selection is performed using sequential forward selection. The next fitted feature is selected if it improves the classification accuracy and it is considered positively correlated with the preceding features. Otherwise, if the feature reduces the accuracy or maintains it constant, it is considered negatively correlated or neutral and is not selected. This selection is repeated until finding the feature that satisfy the group criteria whether nearest or furthest and improve the classification accuracy. This process is repeated for selecting each subsequent feature. The stopping criteria is either obtaining the subset that achieves the best possible accuracy or reaching the maximum allowed number of features for the different ants. By the end of generation, the pheromone values of the features that are part of the best solution are updated.

After that, a new generation is started with the updated pheromone values for features to generate different features subsets in the next generation. By the end of all generations, the features subsets that give the best accuracy in all generations and by different ants for each group are obtained, and then the best global subsets by different groups of ants are obtained. Figure 1 illustrates the full process of feature selection using the proposed ACO model.

4.1 Class-separability

As mentioned previously, the heuristic value is computed using the class-separability approach. Class-separability (CS) [30] is an approach used for feature selection. CS of feature i is defined as:

$$CS_i = SB_i / SW_i \quad (6)$$

Where

$$SB_i = \sum_{k=1}^{K} (\bar{x}_{ik} - \bar{x}_i)^2 \quad (7)$$

$$SW_i = \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \quad (8)$$

$$\bar{x}_{ik} = \sum_{j \in C_k} x_{ij} / n_k \quad (9)$$

$$\bar{x}_i = \sum_{j=1}^{n} x_{ij} / n \quad (10)$$

SB_i is the sum of squares of between class distances (the distances between samples of different classes). SW_i is the sum of squares of within class distances (the distances of samples within the same class). In the whole data set, there are K classes. C_k refers to class k that includes n_k samples. x_{ij} is the value of feature i in sample j. $\bar{x}_{ik}$ is the mean value in class k for feature i. n is the total number of samples. $\bar{x}_i$ is the general mean value for feature i. A CS is calculated for each feature. A larger CS indicates a larger ratio of the distances between different classes to the distances within one specific class. Therefore, CS can be used to measure the capability of features to separate different classes.

4.2 K-Nearest Neighbor

KNN (also known as Instance-based Learning) is a simple efficient data mining technique used for classifying data points based on their distance to other points in a training dataset. KNN is a lazy learner where the training data is loaded into the model and when a new instance need to be classified it looks for the specified k number of nearest neighbors; then, takes a vote to see where, the instance should be classified. For example, if k is 7, then the classes of 7 nearest neighbors are detected. KNN depends on a simple principle which is "similar instances have similar class labels". Distance functions are used to measure similarity between samples. KNN calculates the distance between the unknown data point and every known data point [31, 32]. The common distance between two data points $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_n)$ is defined as follows:

$$Dist(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \quad (11)$$

Euclidian distance is the easiest and most common distance calculation function for quantitative data where $p = 2$.

```
Initialize parameters [number of generations, number of groups of ants, number of ants,
max number of features, pheromone value (τ), heuristic value(η), pheromone evaporation
rate(ρ),local pheromone update parameter(Z), α, β]

While Generation number not exceeded
   For each group of ants
      For each ant
         o Select a start distinct feature randomly
         o Calculate the classification accuracy for the initially randomly selected
           feature
      End for
   End for

   For each ant in each group of ants
      While assigned number of features not exceeded (max number of features) and
      accuracy can still be improved
         o Calculate the fitness value for each feature that can be selected using
           eq.(1)
         o Each group of ants select the candidate feature using different criteria
         o Calculate the accuracy for the subset obtained with the new added feature
         o If the calculated accuracy less than or equal the previous accuracy select
           consecutive feature sequentially with next appropriate fitness value
         o If the maximum accuracy is achieved or max number of feature is reached;
           Final set of features is obtained for the current ant
           Else update the pheromone value of the selected feature locally by eq.(4)
           End If
      End while
   End for

   o Find the features subsets that achieved the best solution in the current generation
     for different groups
   o Update the pheromone values of the features that are part of the best solution
     using eq.(5)to increase these features selection opportunity in next generations
End while
Obtain the global best collected subsets of features from all different groups of ants
that achieved best possible accuracy in all generations
```

Pseudo code 1: The proposed ant colony algorithm for feature selection.

5 Experimental results and discussion

This section shows an empirical performance evaluation of the proposed ACO model. In order to evaluate the proposed model, real world medium-sized datasets shown in Table 1 are tested. The used datasets are heart disease, breast cancer and thyroid which contain 13, 30, 21 features and 303, 569, 7200 samples respectively. The samples are randomly divided into 257/46, 423/146 and 6000/1200 for training and testing respectively. Matlab® 2015a software on an Intel®Core™ i7 CPU @ 1.6 GHz computers is used for implementation. Extensive experimental studies had been tried in order to get the best features subsets selected by ACO which give the highest possible accuracy using KNN classifier. The values of the parameters have been tuned in the experiments as shown in Table 2.

Different features subsets have been tried starting from 2 features subsets until the maximum number of features specified by the user is reached or the best possible accuracy is achieved. In each trial, two different groups of ants, each consists of 3 ants, start to select the features. A number of 100 iterations or generations are executed to reach the highest possible accuracy. Each of the two groups of ants uses a different criterion to select the features. The first group uses the nearest feature to the previously selected one according to the fitness value. The other group uses the furthest feature to the previously selected one according to the fitness value. The pheromone of the selected feature has been increased by small value using eq. (4). At the end of each iteration, the pheromone value of the features that are part of the best solution of either group is incremented by a significant value calculated using eq. (5).

Table 3 illustrates the idea of the proposed ACO model for features selection of breast cancer dataset using two ant groups; nearest and furthest. For illustration purpose, the table includes samples of features subsets with the accuracy which had been recorded as the best accuracy in different generations to clarify the idea. For example, if by the end of a generation, the best achieved accuracy was 93.84% using the two groups of ants. The nearest group achieved it using the features subset {8, 29} and the furthest group achieved it using {24, 4} subset, so the pheromone value of both group subsets will be updated for that generation. The pheromone value of these features will be increased using eq. (5) to have a chance to be part of the best solution in their groups in the next generation.

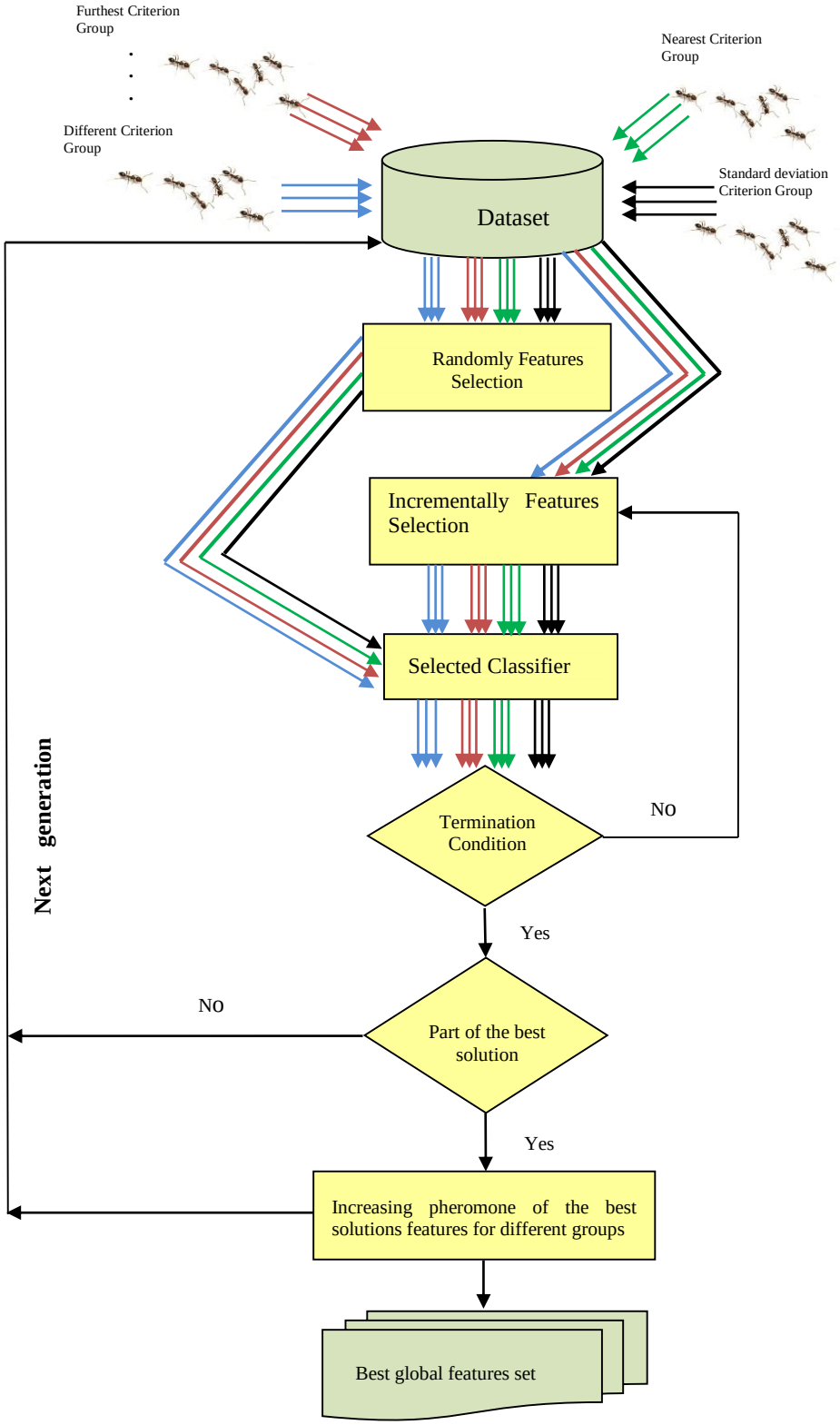


Figure 1: The full process of feature selection using the proposed ACO approach.

Table 1: The used datasets.

Data set name	No. of features	No. of samples	No. of classes	Citation
Heart Disease	13	303	5	[33]
Breast cancer (Wisconsin diagnostic)	30	569	2	[34]
Thyroid	21	7200	3	[35]

In many cases, the same subset can be reached by the two groups either in the same generation or through different generations; this may be due to the limitation of selecting positively correlated features. Also the nearest or the furthest features may change with the generations. This is due to the local changes of the pheromone's values of the selected feature and global changes of the best features by the end of generation which causes the changing in the distances between features. As an example, the 2 subsets that achieved 94.52 % accuracy which are {8,29,30; 8,29,6}. In the first set the furthest feature to the feature 29 was 30 and in the second set it was 6.

Table 2: Parameters of ACO.

Parameters	Values
α	1
β	0.9
ρ	0.9
Z	0.4
No. of generations	100
groups of ants (G)	2
no. of ants/G (na)	3

The best global features subset in the breast cancer dataset that achieved the best possible accuracy which is 97.95% with the least number of features which is 4 features is {16, 28, 7, 1}. By increasing the number of features above 4, the accuracy remains constant as shown in Table 4. So, the selected 4 features are considered the best combination of features that satisfy the best possible accuracy. Figure 2 present the relation between the accuracy versus number of features for breast cancer dataset. The selected features that achieved the best possible accuracy which is 97.95% are the features of cell nucleus mentioned below, these features take decimal values:

1 - mean radius (mean of distances from center to points on the perimeter)

7 - mean of concavity (severity of concave portions of the contour)

16 - standard error of compactness (perimeter^2/area-1.0)

28 - worst concave points (number of concave portions of the contour)

Regarding the heart disease dataset, as shown from Table 5, although the best possible accuracy which is 96.88% has been achieved by 4 features, the best possible accuracy using 2 features is close to it which is 96.77% and it needs only 2 features. So, the set {9, 12} is recommended as best solution. With increasing the features set above 4 features, the accuracy decreases as shown in Table 5. The best selected features are "number of major vessels colored by fluoroscopy" and "exercise induced angina". The first feature takes a value from 0 to 3, the second feature is Boolean and takes values (1 = yes; 0 = no). Figure 3 shows the relation between the accuracy versus number of features for heart disease dataset.

For the thyroid dataset, as shown in Table 6, although the best possible accuracy which is 98.5% has been achieved by 6 features, the best possible accuracy using 4 & 5 features is close to it which is 98.25%, 98.33% respectively. With increasing the features set to 7 features the accuracy remains constant. Figure 4 shows the relation between the accuracy versus number of features for thyroid dataset. The best selected features are TSH (real [0.0 - 0.53]), thyroid surgery (integer [0, 1]), on thyroxin (integer [0, 1]), and FTI (real [0.0020 - 0.642]).

Table 7 shows the percentage of features reduction and the achieved accuracy before and after features reduction for different datasets. Figure 5 presents the total number of features and the reduced number of features using the proposed ACO model for different datasets. Figure 6 presents the comparison of the accuracy for the total number of features and the reduced number of features for the three datasets used. It is clearly that the selected features achieve higher accuracy than the total number of features which ensure that noisy and irrelevant features mislead the classifiers.

Table 8 shows the comparison of the proposed model with the previous work. Since the main purpose of features selection is to achieve the highest accuracy with the least number of features, a comparison between the performance of the proposed research and the previous closest work, will be limited to those that used the same dataset to simplify the comparison capability. By comparing the proposed model with previous work, it seems that the proposed model outperforms others with even less number of features for all databases. Except for breast cancer, Wang G., et al. [21] achieved 98.12% accuracy which is a bit better than ours (97.95%), but with larger numbers of features which are13.5 features rather than only 4 features with our suggested algorithm.

To investigate the capability of the proposed model to achieve promising results with large dataset, the SRBCT microarray dataset [36] which contains 2308 features (genes) and 83 samples was used. The samples are divided into 63/20 for training and testing respectively. It achieved 100% accuracy with 4 genes only as shown in Table 9, the percentage of features reduction is 99.82%. After applying the proposed model on SRBCT dataset, it is concluded that it also has the ability to select features subset which

Table 3: samples of selected features subsets and achieved accuracy using ACO model through different generations for breast cancer dataset.

No. of feature	nearest	Best Acc. %	furthest	Best Acc. %
2	{8,29; 24,4}	93.84	{24,4; 8,29; 6,16; 23,1}	93.84
2	{18,28; 1,21}	94.52	{2,23; 18,28; 21,1}	94.52
2	{4,23; 26,1}	95.21	{4,23}	95.21
3	{ 27,3,23; 8,29,30; 8,29,9; 26,21,3}	94.52	{8,29,30; 8,29,6; 20,23,2; 1,23,12; 18,23,2;16,23,2}	94.52
3	{ 28,18,30; 23,3,14; 21,12,1}	95.89	{14,3,23; 3,23,14}	95.89
3	{ 1,28,7}	97.26	{7,28,1}	97.26
4	{12,21,3,1; 3,23,14,2; 1,21,29,12}	96.58	{ 21,12,1,13; 8,29,19,10; 12,20,27,3}	96.58
4	{7,28,1,16}	97.95	{16,28,7,1}	97.95
5	{28,8,7,1,17; 5,7,28,1,17}	97.26	-	97.26
5	-	97.95	{10,28,7,1,16; 30,28,7,1,16}	97.95
6	{18,7,16,20,28,1;28,16,7,18,10,1}	97.95	{19,28,7,1,17,30}	97.95
7	{25,7,28,11,17,16,10;18,29,6,21,1,12,27}	97.26	-	-
7	-	-	{19,28,10,7,17,1,30}	97.95

Table 4: The selected features subsets using the proposed ACO model that achieved the best accuracy/feature subset for breast cancer.

No. of selected features	**The reduced features subsets**	**Best Accuracy (%)**
2	{4,23; 26,1}	95.21
3	{7,28,1}	97.26
4	{16,28,7,1}	97.95
5	{10,28,7,1,16; 30,28,7,1,16; 15,28,7,1,16; 7,28,30,1,17}	97.95
6	{18,7,16,20,28,1; 28,16,7,18,10,1; 19,28,7,1,17,30}	97.95
7	{19,28,10,7,17,1,30}	97.95
Most important features subset	Features of cell nucleus: 1 - mean radius 7 - mean of concavity 16 - standard error of compactness 28 - worst concave points	

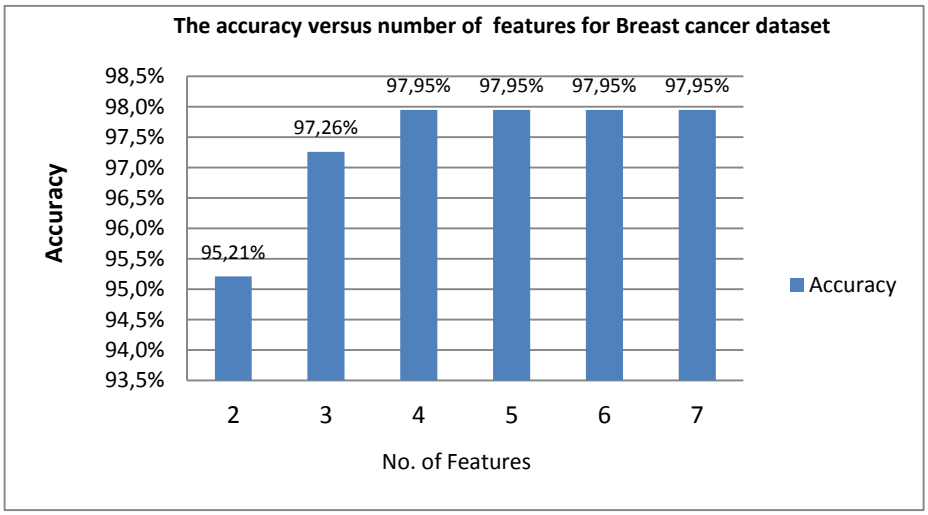


Figure 2: The accuracy versus number of features for breast cancer dataset.

achieve the highest accuracy from large number of features.

Computational Complexity

The actual computational cost of an algorithm can be determined by investigating the computational complexity according to the form of big-O notation. Meta-heuristic algorithms are simple in terms of complexity, and thus they are easy to implement. ACO algorithm has three inner loops n the number of groups of ants; na is the number of ants; and f is the number of selected features, and one outer loop t for iteration. So, the time complexity is $O(n * na * f * t)$. In the experimental studies the inner loops are small (n = 2; na =3; F = 2-7) and (t =100), so the computational cost is relatively inexpensive. The main computational cost will be in five steps according to Pseudo code1: (i)

Table 5: The selected features subsets using the proposed ACO model that achieved the best accuracy /feature subset for heart disease.

No. of selected features	The reduced features subsets	Best Accuracy (%)
2	{9,12}	96.77
3	{6,12,9; 4,12,1}	93.75
4	{4,12,1,6}	96.88
5	{4,12,6,10,2}	94.12
6	{4,6,5,7,1,12}	93.1
7	{4,6,5,7,3,8,12; 4,13,2,10,3,8,12}	84.62
Most important features subset	12 - ca: number of major vessels (0-3) colored by fluoroscopy 9 - exang: exercise induced angina (1 = yes; 0 = no)	

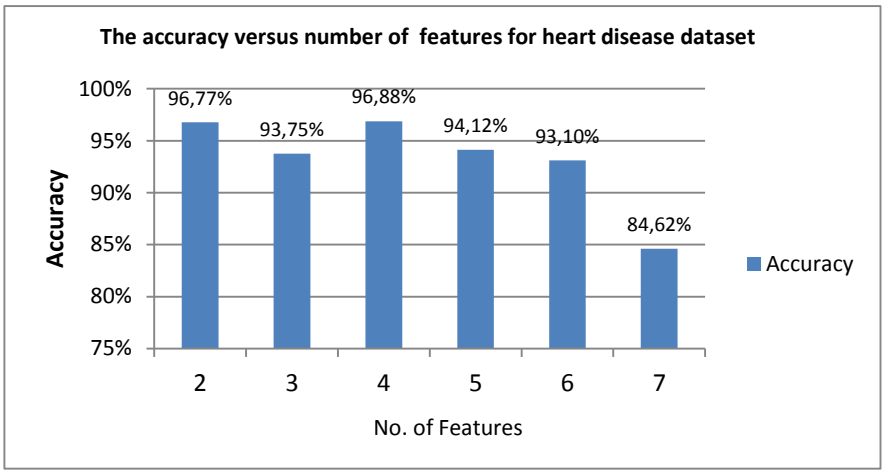


Figure 3: The accuracy versus number of features for heart disease dataset.

Table 6: The selected features subsets using the proposed ACO model that achieved the best accuracy/feature subset for Thyroid.

No. of selected features	The reduced features subsets	Best Accuracy (%)
2	{3,17}	97
3	{21,17,3}	97.92
4	{8,17,3,21}	98.25
5	{19,17,5,3,8; 8,17,3,21,16; 3,17,6,21,8}	98.33
6	{21,17,9,3,16,8}	98.5
7	{3,17,12,21,9,16,8; 21,17,5,3,9,8,16}	98.5
Most important features subset	17 - TSH 8 - Thyroid_surgery 3 - On_thyroxine 21 - FTI	

random feature selection, (ii) subset selection using different criteria, (iii) updating pheromone values (iv) calculating probabilistic transition rule, and (v) termination condition.

For example, the estimated time for heart disease to select 2 features was ≅ 15.3 sec on average, selecting 3 features needs 35.5 sec on average and selecting 4 features needs 48 sec on average.

6 Conclusion and future work

In this research, an ACO model has been developed to select minimum subsets of features from datasets that can achieve the best possible accuracy. The purpose was to reduce redundant, irrelevant and noisy features that mislead the classifier. The proposed model use different groups of ants to select different features subsets that give the best possible result. Each group uses a different criterion to select the features. In this research two different criteria have been applied; the nearest and furthest criteria. By the end of each generation, each group selects the best features subsets that achieve the best accuracy for that group. The pheromone values of these features are increased to be given higher opportunity to be part of the selected features in the next generation for that group. By the end of all generations, the best features subsets for each group have been selected, and then the global best solutions from all groups have been reached.

The results showed that, the right selection of features and eliminating irrelevant, noisy and redundant features increase the accuracy of classifiers. The percentage of

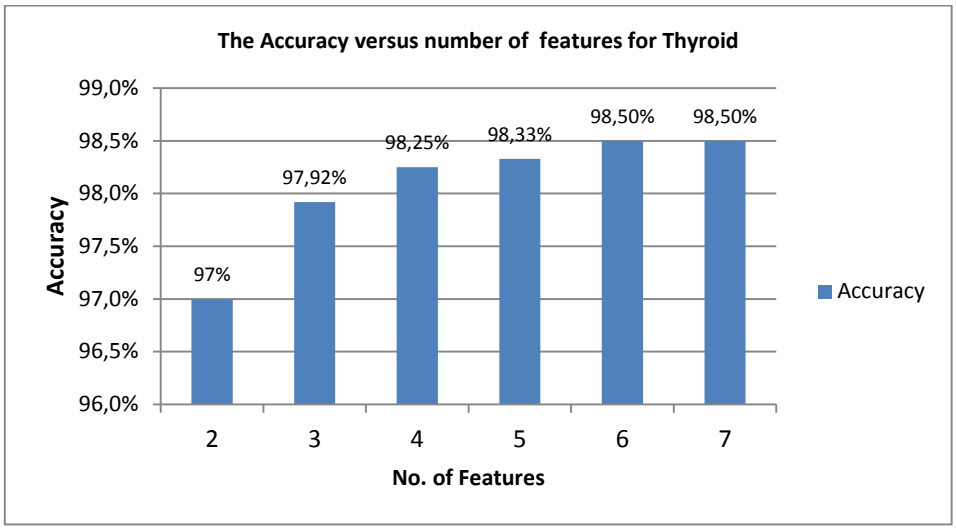


Figure 4: The accuracy versus number of features for Thyroid dataset.

Table 7: The percentage of features reduction and the achieved accuracy for different datasets.

Data sets	Training size/Testing	No. of features	Reduced subset	% Reduction	Accuracy of total features	Accuracy of reduced features
Heart disease	257/46	13	2	84.61 %	82.5%	96.77%
Breast cancer (Wisconsin diagnostic)	423/146	30	4	86.66 %	93.15 %	97.95 %
Thyroid	6000/1200	21	4	80.95 %	93.15%	98.25 %

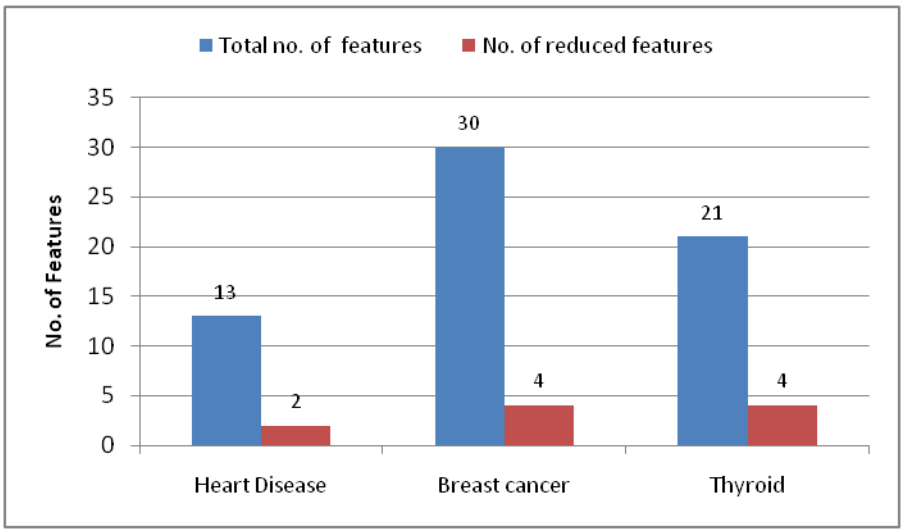


Figure 5: The total no. of features compared with reduced no. of features for different datasets.

features reductions are 84.61 %, 86.66 %, and 80.95 % for heart disease, breast cancer, and thyroid respectively with that the accuracy had increased from 82.5% to 96.77%, from 93.15 % to 97.95 % and from 93.15% to 98.25 % respectively. By trying the model on different datasets it achieved promising results compared with previous works, it achieved higher accuracy with less number of features for all databases. Different features subsets have been reached using different groups' criteria which give the capability to collect different solutions and reach the best global solutions. The proposed model proved its capability to select features from large datasets, when applied on SRBCT microarray dataset. As a future work, other criteria can be used to collect different subsets of features, also parametric studies can be studied and applying the model on different datasets can be done.

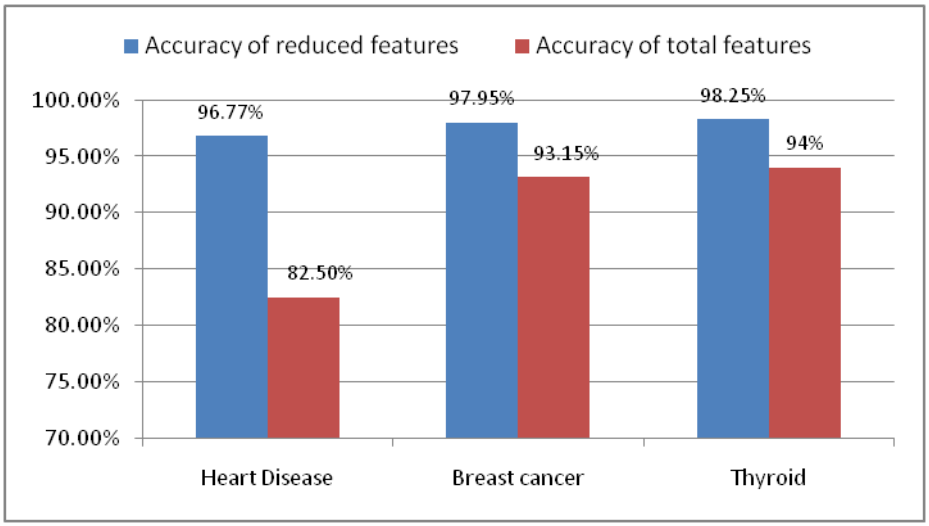


Figure 6: The comparison of the accuracy for reduced no. of features and the total no. of features.

Table 8: The comparison of the proposed model with the previous work.

Data sets	No. of features	No. of Reduced Features/ Accuracy				
		Proposed system	Sivagaminathan, et al. [16]	Dadaneh, et al. [20]	Wang G., et al. [21]	Chen, et al. [23]
Heart Disease	13	2 / 96.77 %	-	-	6.1/88.22% On average	8.08/ 86.67% On average
Breast cancer (Wisconsin diagnostic)	30	4 / 97.95 %	12 / 95.57 %	11 / 91.23%	13.5/98.12% On average	5.89 / 95.99 On average
Thyroid	21	4 / 98.25 %	14 / 94.5 %	-	-	-

Table 9: The selected features subsets using the proposed ACO model that achieved the best accuracy /feature subset for SRBCT.

No. of selected features	The reduced features subsets	Best Accuracy (%)
2	{1613, 1165; 1427, 1479}	85
3	{156, 1606, 1601}	95
4	{1434, 1775, 2214, 1920}	100
Most important features subset	1434- kinesin family member 3C 1775- guanosine monophosphate reductase 2214- FYN oncogene related to SRC, FGR, YES 1920- lamin B2	

7 References

[1] S. M. Vieira, J. M. Sousa, and T. A. Runkler, (2009). Multi-criteria ant feature selection using fuzzy classifiers. In Swarm Intelligence for Multi-objective Problems in Data Mining, Springer Berlin Heidelberg, 19-36.

[2] I. A. Gheyas, L. S. Smith, (2010). Feature subset selection in large dimensionality domains. Pattern Recognition 43(1) 5-13.

[3] A. Unler, A. Murat, (2010). A discrete particle swarm optimization method for feature selection in binary classification problems. Eur. J. Oper. Res. 206(3) 528–539

[4] M. Dash, K. Choi, P. Scheuermann, and H. Liu, (2002). Feature selection for clustering filter solution. In: Proc. of Second International Conference on Data Mining ICDM 115–122.

[5] P. Mitra, C. A. Murthy, and S. K. Pal, (2002). Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Analysis and Machine Intelligence 24(3) 301–312

[6] A. Miller, (2002). Subset Selection in Regression. (2nd ed.). Chapman & Hall/CRC, Boca Raton.

[7] Blum, L. Avrim and P. Langley, (1997). Selection of relevant features and examples in machine learning. Artificial intelligence, 1, 245-271.

[8] L. Talavera, (2005). An evaluation of filter and wrapper methods for feature selection in categorical clustering. In International Symposium on Intelligent Data Analysis, Springer Berlin Heidelberg, 440-451.

[9] L. A. M. Pereira, D. Rodrigues, T. N. S. Almeida, C. C. O. Ramos, A. N. Souza, X-S. Yang, and J. P. Papa, (2014). A Binary Cuckoo Search and Its Application for Feature Selection, In Cuckoo Search and Firefly Algorithm. Springer International Publishing 141-154.

[10] S. Fong, S. Deb, X. S. Yang, and J. Li, (2014). Feature selection in life science classification: metaheuristic swarm search. IT Professional 16(4) 24-29.

[11] K. H. Chen, L. F. Chen, and C. T. Su, (2014). A new particle swarm feature selection method for classification. Journal of Intelligent Information Systems 42(3) 507-530.

[12] Thanh Tung Khuat, My Hanh Le, (2016).Optimizing Parameters of Software Effort Estimation Models using Directed Artificial Bee Colony Algorithm, Informatica 40, 427–436.

[13] B. Xue, M. Zhang, and W. N. Browne, (2013). Particle swarm optimization for feature selection in classification: a multi-objective approach, IEEE transactions on cybernetics 43(6) 1656-1671.

[14] A. Khazaee, (2013). Heart beat classification using particle swarm optimization, International Journal of Intelligent Systems and Applications 5(6) 25.

[15] W. C. Yeh, Novel swarm optimization for mining classification rules on thyroid gland data, Information Sciences 197 (2012) 65-76.

[16] R. K. Sivagaminathan, and S. Ramakrishnan, (2007). A hybrid approach for feature subset selection using neural networks and ant colony optimization. Expert systems with applications 33(1), 49-60.

[17] J. B. Jona, and N. Nagaveni, (2014). Ant-cuckoo colony optimization for feature selection in digital mammogram. Pakistan Journal of Biological Sciences 17(2), 266.

[18] A. Asad, A. T. Azar, N. El-Bendary, and A. E. Hassaanien, (2014). Ant colony based feature selection heuristics for retinal vessel segmentation. arXiv preprint arXiv:1403.1735.

[19] A. J. Tallon-Ballesteros and J. C. Riquelme, (2014). Tackling Ant Colony Optimization Meta-Heuristic as Search Method in Feature Subset Selection Based on Correlation or Consistency Measures. in International Conference on Intelligent Data Engineering and Automated Learning. 2014. Springer, 386–393.

[20] Behrouz Zamani Dadaneh, Hossein Yeganeh Markid, Ali Zakerolhosseini, (2016). Unsupervised probabilistic feature selection using ant colony optimization. Expert Systems with Applications, 5327–42.

[21] G. Wang, H. E. Chu, Y. Zhang, H. Chen, W.Hu, Y. Li, and X. J. Peng, (2015). Multiple parameter control for ant colony optimization applied to feature selection problem. Neural Computing and Applications 26(7), 1693-1708.

[22] Bo Liu, Mei Cai and Jiazong Yu, (2015). Swarm Intelligence and its Application in Abnormal Data Detection. Informatica 39(1), 63–69.

[23] B. Chen, L. Chen, and Y. Chen, (2013). Efficient ant colony optimization for image feature selection. Signal processing 93(6) 1566-1576.

[24] C. Coello, S. Dehuri, and S. Ghosh, eds, (2009). Swarm intelligence for multi-objective problems in data mining. 242 Springer.

[25] Kanan, H. Rashidy, K. Faez, and S. M. Taheri, (2007). Feature selection using ant colony optimization (ACO): a new method and comparative study in the application of face recognition system. In Industrial Conference on Data Mining, Springer Berlin Heidelberg, 63-76.

[26] M. Dorigo, (1992). Optimization, learning and natural algorithms, Ph.D. Thesis, Politecnico di Milano, Italy.

[27] M. Dorigo, V. Maniezzo, A. Colorni, A., (1996). Ant system: optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 26(1), 29-41.

[28] S. Tabakhi, M. Parham, and F. Akhlaghian, (2014). An unsupervised feature selection algorithm based on ant colony optimization. Engineering Applications of Artificial Intelligence 32, 112-123.

[29] S. Dehuri, S. Ghosh, and C. A. Coello, (2009). An introduction to swarm intelligence for multi-objective problems. In Swarm Intelligence for Multi-objective Problems in Data Mining, Springer Berlin Heidelberg, 1-17.

[30] Feng chu & lipo wang, (2005). Applications of support vector machines to cancer classification with microarray data. International Journal of Neural Systems, 15(6), 475–484

[31] C. C. Aggarwal, 2015. Data Mining: The Textbook, Springer International Publishing Switzerland.

[32] M. Kubat, 2015. An Introduction to Machine Learning, Springer International Publishing Switzerland.

[33] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[34] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[35] http://sci2s.ugr.es/keel/category.php?cat=clas

[36] Alexander Statnikov, C.F.A., Ioannis Tsamardinos. Gene Expression Model Selector. 2005 [cited 2017 April]; Available from: www.gems-system.org

Landmarking-Based Unsupervised Clustering of Human Faces Manifesting Labio-Schisis Dysmorphisms

Daniele Conti
Department of Applied Science and Technology, Politecnico di Torino
corso Duca degli Abruzzi 24, 10129 Torino, Italy

Luca Bonacina
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Antonio Froio
Department of Energy, Politecnico di Torino, corso Duca degli Abruzzi 24, 10129 Torino, Italy

Federica Marcolin and Enrico Vezzetti
Department of Management and Production Engineering, Politecnico di Torino
corso Duca degli Abruzzi 24, 10129 Torino, Italy
E-mail: federica.marcolin@polito.it

Domenico Speranza
Dipartimento di Ingegneria Civile e Meccanica, Universit degli Studi di Cassino e del Lazio Meridionale
Viale dell'Universit, 03043 Cassino (Frosinone), Italy

Keywords: facial dysmorphism, labio-schisis, diagnosis, feature extraction, landmarking, clustering, D-MST, artificial intelligence, decision support

Received: July 20, 2016

Ultrasound scans, Computed Axial Tomography, Magnetic Resonance Imaging are only few examples of medical imaging tools boosting physicians in diagnosing a wide range of pathologies. Anyway, no standard methodology has been defined yet to extensively exploit them and current diagnoses procedures are still carried out mainly relying on physician's experience. Although the human contribution is always fundamental, it is self-evident that an automatic procedure for image analysis would allow a more rapid and effective identification of dysmorphisms. Moving toward this purpose, in this work we address the problem of feature extraction devoted to the detection of specific diseases involving facial dysmorphisms. In particular, a bounded Depth Minimum Steiner Trees (D-MST) clustering algorithm is presented for discriminating groups of individuals relying on the manifestation/absence of the labio-schisis pathology, commonly called cleft lip. The analysis of three-dimensional facial surfaces via Differential Geometry is adopted to extract landmarks. The extracted geometrical information is furthermore elaborated to feed the unsupervised clustering algorithm and produce the classification. The clustering returns the probability of being affected by the pathology, allowing physicians to focus their attention on risky individuals for further analysis.

Povzetek: Predstavljena je D-MST metoda za nenadzorovano grupiranje slik obrazov za diagnosticiranje.

1 Introduction

Medical imaging has seen an important enhancement in the past decades thanks to various technological achievements. Magnetic Resonance Imaging (MRI), Computed Axial Tomography (CAT), X-ray imaging, Ultrasound scans imaging (US) provide physicians with valuable information to diagnostic purpose. In particular, foetal diseases attracted attentions and efforts with the common aim to improve the current diagnosis techniques, fostered by the objective of defining a tailored therapy as early as possible. A crucial role in this activity is played by three-dimensional ultrasound scans, which could provide in-depth detailed images of foetal morphology in a safe and non-invasive way. Despite technological improvements, medical image-driven diagnosis suffers the deficiency/absence of automated computer science treatment, even for diseases such as Fetal Alcohol Syndrome (FAS) and labio-schisis [1–5]. This work aims to provide a methodology and a tool for supporting the diagnosis of labio-schisis pathology (cleft lip), which has been chosen due to its relatively large incidence in the population [6]. This task is conceived for prenatal diagnosis and stems from a recently developed work [7], in which an automatic procedure was designed to process a stack of 2D ultrasound scans of foetal faces

by transforming the standard DICOM images in a PLY 3D model. The core of the proposed method relies on the clustering technique. Common algorithms for unsupervised data clustering belong to two main categories: partitioning algorithms and hierarchical clusterings [8]. Algorithms in the first class, such as K-means or the recently proposed Affinity Propagation (AP) [9], define a subset of individuals called *centroids*, i.e. the class exemplars, to which any other node is compared. Hierarchical clustering algorithms, such as Single Linkage, compare couples of individuals and merge the closest in a class, thus creating a chain of hierarchical dependencies. In the first case, the expected number of classes, i.e. of centroids, should be *a priori* defined (except for Affinity Propagation), while in the second case the pruning of the hierarchical tree specifies how many clusters to be returned [10]. Among them, the bounded Depth Minimum Steiner Trees (D-MST) unsupervised clustering algorithm is chosen for this study [11], [12].

For privacy reasons, after the feasibility test on an ideal foetuses dataset, the public Bosphorus database was adopted, containing facial depth maps of 105 adult individuals showing the seven fundamental facial expressions, [13]. The defect was artificially simulated on the faces by modifying some Bosphorus point clouds. This way, seven artificial faces were generated with left-sided and right-sided labio-schisis. The algorithm is designed to be robust against different defect types.

The work is structured as follows. Firstly, an outline of geometrical face description formalization is presented together with related feature extraction aimed at landmarks localization. Then, information coming from geometrical descriptors are exploited to feed the unsupervised D-MST clustering algorithm for discriminating individuals according to the presence/absence of the pathology.

2 Methods

The algorithm is meant to detect the presence/absence of cleft lip in a query face. It is designed to work with three-dimensional foetal faces obtained through automatic elaboration of bidimensional ultrasound scan stacks. On the other hand, it has been extensively tested with a large size adult individuals dataset.

2.1 Mathematical background

Bosphorus database provides coordinates of facial point clouds, obtained through laser scans, as a binary file. A pre-built routine is provided together with data for reading binary files, extract cloud points data, and return information as a matrix containing the Cartesian coordinate of each point. The facial surface can be seen from the mathematical standpoint as the locus defined as

$$z \in \mathbb{R}^3 : z = f(u, v)$$

and it can be referred to as a *free-form surface*. A free-form surface is required to be smooth, with normal vector defined almost everywhere but edges, cusps, etc., but not belonging to a simple mathematical class of surfaces, like conics for example. Anyway, it can be divided in subdomains, each of them treatable as a linear combination of simple geometries. Thus, we define a surface patch divided in domains as an n-tuple of functions:

$$f(u, v) = (f_1(u, v), f_2(u, v), ..., f_n(u, v)). \quad (1)$$

Taking advantage of this definition, in order to objectively compare one face to another, the surface is point-by-point mapped-on with entities belonging to the Differential Geometry domain, here called *geometrical descriptors*. Twelve different geometrical descriptors, together with first and second derivatives, are chosen: three coefficients of the first fundamental form, i.e. E, G, F, three coefficients of the second fundamental form, i.e. e, f, g, the Gaussian curvature K, the mean curvature H, the principal curvatures k_1 and k_2, the shape index S and the curvedness index C. In the following section we go through the adopted geometrical descriptors definitions [14, 20].

2.2 Geometrical descriptors

A free-form surface is not an Euclidean geometry. Thus, distances on a face cannot be computed with the standard formula $s^2 = \sum_{i=0}^{d}(u_i - v_i)^2$. The first fundamental form, also called Riemann metric, allows to define equivalent concept of distance upon a non-Euclidean surface. For $d = 2$, the infinitesimal distance element ds can be defined as $ds^2 = Edu^2 + 2Fdudv + Gdv^2$. E, F, G are the first fundamental form coefficients. They can also be expressed in terms of partial derivatives as

$$E = \|f_u\|^2, \quad (2)$$

$$F = \langle f_u, f_v \rangle, \quad (3)$$

$$G = \|f_v\|^2, \quad (4)$$

where $f_u = \frac{\partial f}{\partial u}$. Moreover, by defining the normal unit vector in point (u, v) belonging to the face domain as

$$N(u, v) = \frac{f_u \times f_v}{|f_u \times f_v|}(u, v), \quad (5)$$

we can also introduce the second fundamental form as $ds^2 = edu^2 + 2fdudv + gdv^2$, with

$$e = \langle N, f_{uu} \rangle, \quad (6)$$

$$f = \langle N, f_{uv} \rangle, \quad (7)$$

$$g = \langle N, f_{vv} \rangle, \quad (8)$$

where $\langle \cdot \rangle$ denotes the scalar product. In order to introduce curvatures, let us consider the tangent plane $T_p(f)$ to f in point $p = f(u_0, v_0)$; it can be defined as the two-dimensional vector subspace $Df(u, v) \subset \mathbb{R}^3$, where D is the functional differential operator. For each point

p, there exists a set of orthonormal vectors $\{e_1, e_2\}$ for the tangent plane T_p, such that $DN_p(e_1) = -k_1 e_1$ and $DN_p(e_2) = -k_2 e_2$, where k_1 and k_2 are called the principal curvatures and e_1 and e_2 the principals directions at p. In terms of the principal curvatures, Gaussian curvature K and mean curvature H can be introduced:

$$K = k_1 k_2 = \frac{eg - f^2}{EG - F^2}, \tag{9}$$

$$H = \frac{k_1 + k_2}{2} = \frac{eG - 2fF + gE}{2(EG - F^2)}. \tag{10}$$

Thus, the principal curvatures can be obtained as the roots of the quadratic equation $k^2 - 2Hk + K = 0$, resulting in

$$k_1 = H + \sqrt{H^2 - K} \tag{11}$$

and

$$k_2 = H - \sqrt{H^2 - K}. \tag{12}$$

An insightful method for evaluating curvatures was introduced by Koenderink and van Doorn [15], who defined the shape index S and the curvedness index C. They can be expressed in terms of the principal curvatures:

$$S = -\frac{2}{\pi} arctan \frac{k_1 + k_2}{k_1 - k_2}, \; with \, S \in [-1, +1], \; k_1 > k_2, \tag{13}$$

$$C = \sqrt{\frac{k_1^2 + k_2^2}{2}}. \tag{14}$$

The range spanned by the shape index can be partitioned into nine different intervals, spanning from cup to dome, each of them representing a particular shape. The curvedness index provides information about how gently the surface bends. Differently from other geometrical descriptors such as the shape index, it is not independent on the unit length and has the dimension of a reciprocal length. These geometrical descriptors are computed for each point of the face and exploited for both landmarking and clustering phases.

2.3 Landmarking

Geometrical descriptors are suitable to be mapped point-by-point on facial surfaces. So, by computing their values for all individuals in the dataset, a distribution of their local behaviour is obtained. Such a statistics can be exploited as characteristic information of the facial region and used to automatically localize facial landmarks. Landmarks are typical facial points, such as the nose tip, i.e. the *pronasal*, the nose basis, i.e. the *subnasal*, the internal and external eye extrema, i.e. the *endocanthions* and *exocanthions*. Figure 1 shows the most renown landmarks.

Landmarks can be automatically detected by setting tailored thresholds, empirically defined, in specific facial areas (where each landmark is more likely to be) for all geometrical descriptor facial maps. Focusing on the specific problem addressed in this work, we developed a program for automatically extracting *pronasal* and *labrum superior* points. For further details about landmarking, please refer to [16–23].

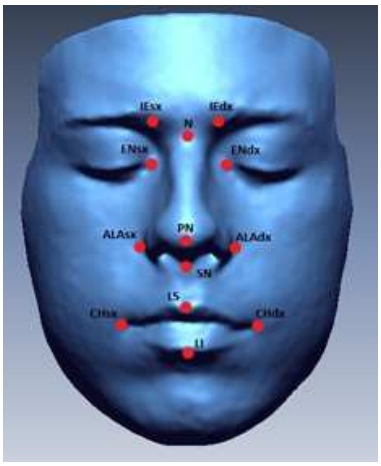


Figure 1: Main soft-tissue landmarks. From top to bottom: IEsx/IEdx, left and right *Inner Eyebrows*. N, *Nasion*. ENsx/ENdx, right and left *Endochantions*. ALAsx/ALArx, *Alae*. PN, *Pronasal*. SN, *Subnasal*. LS, *Labrum superior*. CHsx/CHdx, righty and left *Cheilions*. LI, *Labrum Inferior*.

2.4 Clustering

In order to perform the clustering, the input database is put in the form of a $N \times M$ matrix with a row for each individual to be classified and as many columns as the number of geometrical descriptors to be exploited for facial description purpose. For instance, considering all individuals available, if we report values of all seventeen geometrical descriptors expressed in the *labrum superior* landmark, a 105 x 17 matrix is created. In our case, values of a subset of geometrical descriptors expressed in an arbitrary surface portion, covering the central part of the upper lip and departing on a straight line from the identified LS landmark, are considered, in order to catch sufficient information about the possible presence of labio-schisis pathology. Clustering algorithms require the definition of a dissimilarity measure to be used for one-to-one face comparisons. It leads to the so called *Dissimilarity Matrix*, S, whose entries $s_{i,j}$ with $i, j = 1, ..., N$ are the dissimilarities between any couple of individuals i and j. $s_{i,j}$ can be defined in different ways, depending on the kind of data to be clustered [24]. Spearman's correlation rank distance $\rho_{i,j}$ is chosen as dissimilarity measure, in order to overcome problems related to descriptors, such as the shape and curvedness indexes, lying on different domains and with different measure scales. In particular:

$$\rho_{i,j} = \sum_{m=1}^{M} \frac{(x_{i,m} - \bar{x}_i) \cdot (x_{j,m} - \bar{x}_j)}{\sqrt{\sum_m (x_{i,m} - \bar{x}_i)^2 \cdot \sum_m (x_{j,m} - \bar{x}_j)^2}}, \tag{15}$$

$$s_{i,j} = 1 - \rho_{i,j}. \tag{16}$$

In 15, the usual definition of correlation is given, being ρ the Spearman's rank correlation coefficient. In particular, the variable x is the data rank, instead of the bare data itself as for Pearson correlation coefficient, i and j identify individuals and index m runs over geometrical descriptors values. In the second equation, s is the dissimilarity. The input dataset is treated as a fully connected weighted graph $G(n, e)$, with $n = 1, ..., N$, $e = \{(n_i, n_j)\}$, $i, j = 1, ..., N$, with N individuals and $\frac{N(N-1)}{2}$ weighted edges e, whose weight is the dissimilarity $s_{i,j}$ between the two individuals. A fictitious node, called *root*, is added to this graph and connected to all other nodes by a weight λ, empirically defined. From a Physical viewpoint, λ can be interpreted as the chemical potential of the system, i.e. the cost for adding an individual to the system itself, and it governs the most probable number of clusters returned. On the other hand, a depth parameter D is introduced to drive the final output. It is a constraint representing the maximum observable depth, namely the distance, in terms of nodes, from the root to the external leaves of clusters. Thanks to this additional parameter, D-MST interpolates between the Affinity Propagation algorithm, returning an arbitrary number of spherical clusters with $D = 2$, in which leaves are directly connected to the centroids via edge means with comparable dissimilarities, and the Single Linkage algorithm, in which $D > N$. In order to perform a classification, two variables (d_i, π_i) are assigned to any node and exploited to define an objective function to be minimized for detecting the optimal spanning tree T^* in the graph. The variable $d_i \in [2, N]$, $d_i \leq D$ is the distance, in terms of number of nodes, from the root, and assumes discrete values; variable $\pi_i = j$, $j \in [1, N]$, $j \neq i$ is a pointer tracking the ancestor of node i. Thus, the cost function is:

$$\begin{aligned} E(\{d_i, \pi_i\}_{i=1}^N) = \\ \sum_i s_{i,\pi_i} + \\ + \sum_{i,j \in \partial i} (h_{i,j}(\pi_i, \pi_j, d_i, d_j) + h_{j,i}(\pi_j, \pi_i, d_j, d_i)), \end{aligned} \tag{17}$$

where h_{ij} is defined as

$$h_{i,j} = \begin{cases} 0 & \{\pi_i = j \Rightarrow d_i = d_j + 1\} \\ -\infty & else, \end{cases} \tag{18}$$

and imposes an artificial constraint to the cost function that requires the returned optimum tree to be connected. In this terms, the probability of observing a configuration of variable for the optimum tree is given by the Boltzmann weight

$$P(\{\pi_i, d_i\}) \propto e^{-\beta E(\{\pi_i, d_i\})} \tag{19}$$

and it is maximized by a message passing algorithm described in [25] and [26].

3 Results

3.1 Preliminary analysis

Individuals' faces are reported in Bosphorus database as point clouds pre-ordered on a square grid and with the same orientation, in particular with nose oriented alongside with z-axis, x-axis aligned with chin-forehead line and, consequently, y-axis aligned from cheek to cheek. A first analysis of data contained in Bosphorus database is conducted by inspection, in order to examine the facial points suitable to our purpose. Indeed, referring only to faces with no expression, some facial point clouds showed degradation and low accuracy in shaping the face itself. In particular, it is not unusual to encounter data with a rough mouth surface that has no actual correspondence to the individual's picture accompanying data. Thus, all corrupted data were excluded from further analysis, resulting in an input matrix collecting 74 healthy individuals.
As a preliminary step, all facial point clouds are cropped in size, limiting the region of interest to a squared area. A four-pixels-side mean-filter is then applied in order to reduce noise and smooth surface peaks, where a pixel is intended as the squared surface area wrapping a point of the face. Moreover, the fact of having pre-oriented faces allowed us to avoid a pre-processing step aimed to provide data with a standard orientation. Most of all, it allowed to easily identify the central region of the face, by looking at those points with relatively higher z-coordinate. This way, the facial area containing nose and a mouth portion is identified and exploited for further analysis.

3.2 Computing geometrical descriptors and landmarking

Geometrical Descriptors are point-by-point computed, starting from derivatives. They are obtained by computing the surface gradients, along x and y directions, then by averaging values obtained in a ten-pixels-side window centred into the point of interest. All other geometrical descriptors can be obtained, as previously shown in Methods section, starting from first and second derivatives, and are easily computed for each point of the facial surface.

Once geometrical descriptors are obtained, they are exploited to define where the *pronasal* landmark and the *labrum superior* landmark are placed upon the surface. Our attention is focused most of all on the latter landmark, as it would affect the chosen area of investigation for the cleft

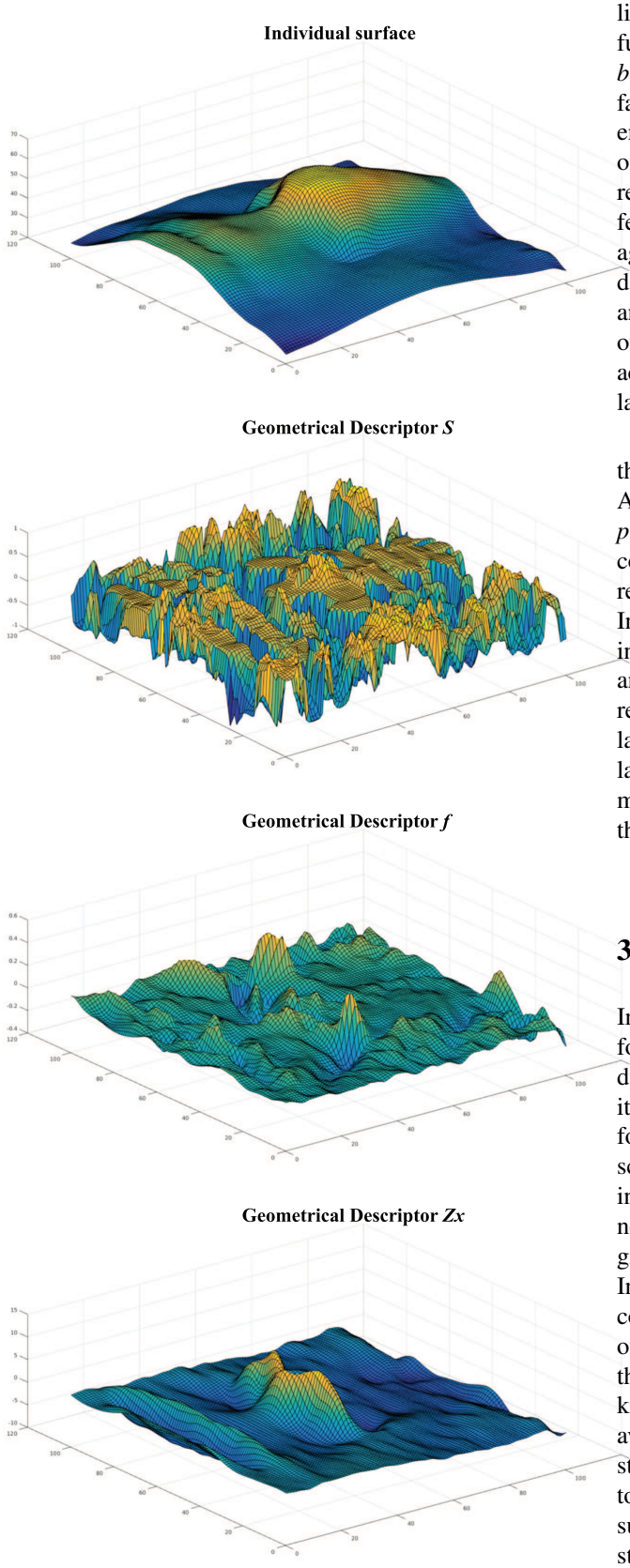


Figure 2: Geometrical descriptors mapped on a face. From top to bottom: bare facial surface, shape index S, second fundamental form f and derivative along x.

lip dysmorphism clustering. Indeed, the *pronasal* is helpful in proceeding with an accurate identification of the *labrum superior* itself. Starting from the central area of the face identified previously relying on z-coordinate, we set empirical constraints to values assumed by meaningful geometrical descriptors. They are the descriptors that, in the region of interest, present a characteristic behaviour. Referring to the previous work [19] and assessing the choice against this database, the shape index S, the second fundamental form coefficient f, and first derivatives along x and y directions have been chosen (figure 2). This subset of descriptors, conveniently constrained, leads to a 100% accuracy in the automatic determination of the *pronasal* landmark for the pre-processed database in exam.

Once the *pronasal* is identified, it is adopted to delimit the region of interest for the *labrum superior* detection. Approximatively half of the area going from chin to the *pronasal* itself is taken into consideration to detect this second landmark. In this region, the previous procedure is repeated changing only the geometrical descriptor adopted. In such a case, the chosen information relies upon the shape index S, the mean curvature H, the first derivative along x, and the second derivative along x. In this case, the accuracy reached by the algorithm is around 94%, but even when the landmark obtained by the algorithm and the ground truth landmark do not match perfectly, their relative distance remains around a few pixels. Thus, the error is not affecting the final output of the algorithm.

3.3 Prenatal applicability

In its original intention, the present work has been designed for pre-birth diagnosis of rare diseases manifesting facial dysmorphisms. Labio-schisis presents high incidence and it is clearly detectable through ultrasound-scans when the foetus is affected. Another connected pathology is palatoschisis. It is more difficult to be observed by US-scans inspection and its current diagnosis techniques would benefit of an automatic procedure for highlighting morphological differences that are symptomatic of the disease itself. In this perspective, the clustering procedure here proposed could be efficient only if it relies on an effective detection of the foetus's facial landmarks, which are clearly fuzzier than those of an adult. In our work, we tested the landmarking algorithm on the limited amount of real foetus data available and found that, after a tiny rearrangement of constraints imposed to the same subset of geometrical descriptors, the *pronasal* and the *labrum superior* landmarks were successfully detected (figure 3). Although this result is not statistically significant, it moves towards the application of such kind of procedure to pre-birth diagnosis. Indeed, once landmarks are identified correctly, the cleft lip manifesting face morphology reports similar differential geometry properties and, thinking toward a clustering perspective, it is totally equivalent to the case of adult individuals.

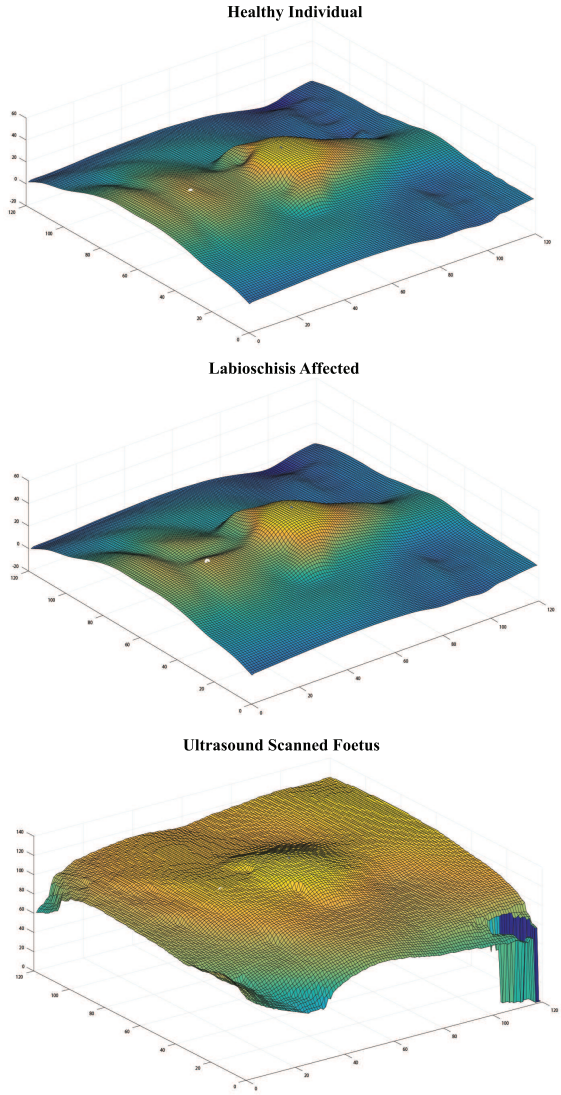


Figure 3: Surface Landmarking. For each surface, the white dot indicates the position of the *labrum superior* landmark. The comparison between faces not manifesting (top) and manifesting (centre) the labio-schisis is presented. Moreover, it is shown the case of landmarking on a foetal 3D model, (bottom). 0-valued points in the last figure indicate missing data.

3.4 Unsupervised clustering

Once the *labrum superior* landmark has been detected, the area of interest for the labio-schisis pathology is identified. In particular, a straight line laying on the upper lip is taken as a biometric information of the individual. The line length and width can be arbitrary, provided that it spans most of the lip itself, without invading other face regions. Therefore, if the cleft lip is present in the individual under investigation, it will also be in the region in exam. Moreover, the pre-orientation of individuals' face simplifies the identification of the lip line, avoiding a useless detection of its complete morphology.

A subset of meaningful geometrical descriptors expressed in any point composing this line is then stored into a row vector, building a matrix collecting all individuals in exam. The opposite in sign of the shape index S and of the coefficient of the first fundamental form G, the first and second derivatives along y of the free-form surface are sufficient to obtain the convergence of the clustering algorithm toward a successful classification. In our specific study, the chosen line is forty pixels-long and its width is three pixels, pinched at the *labrum superior* landmark. This choice is useful for the application of a median-filter on any three-pixels-side square of geometrical values spanning the line. Median filter smooths the descriptor behaviour along the surface, without adding artificial information to the one extracted from the geometrical analysis of the surface. In the end, the input matrix M presents a number of rows equal to 74 (number of healthy faces) plus 7 (number of artificially-induced cleft lip-affected faces) and a number of columns equal to 40 times the number of chosen geometrical descriptors.

Starting from M, the so called *similarity matrix* is computed. As previously mentioned, the similarity matrix S is a squared symmetric matrix which reports, with any of its entries s_{ij}, a measure of distance between any couple of individuals. In particular, the most suitable choice for the kind of data handled in this work is the Spearman's correlation rank distance, computed as $s_{ij} = 1 - \rho_{ij}$, where ρ is shown explicitly in equation 15. Once the similarity matrix is computed and the specific clustering depth D is set, the unsupervised clustering D-MST can be run.

As specified previously in the section Methods, D-MST is governed by an external parameter λ influencing the number of identified clusters. Lower values of λ would lead toward many clusters and non-linked individuals, while larger values of the parameter would return a single cluster collecting all nodes of the graph. In general, the maximum value of dissimilarities s_{ij} found in the S matrix is taken as upper bound for the parameter. In order to identify the most proper value of lambda to be chosen, stability regions of the clustering algorithm are investigated. They are intended as regions of the parameter space in which the algorithm converges toward a stable solution, in terms of number of clusters, despite the parameter change. So, the range between the minimum value of dissimilarity, exclu-

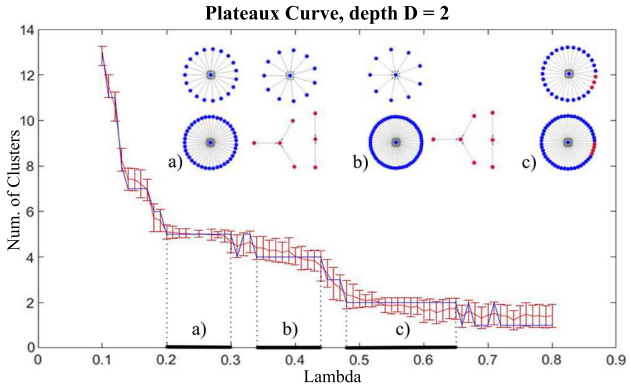


Figure 4: Number of clusters versus λ, for $D = 2$. Plateaux curve. The higher the λ value, the lower the number of clusters. For the plateaux associated to 5, 4 and 2 clusters, the relative clustering are overlaid. In particular, observing plateau a), two cleft lip clusters are detected correctly, even separating left- and right-sided cleft lips; on the contrary, healthy individuals population is divided in subpopulations as well. Increasing λ, blue clusters do not form a single class, while individuals affected by the pathology are merged to healthy clusters.

ded zero, and the maximum one, is linearly split up into a large number of bins. For each of them, the algorithm is run fifty times, in order to return results with a statistical significance. In such a way, what here is called *plateaux curve* is built and it plots the number of clusters returned as a function of the value of λ.

Such a procedure is repeated for the different depths investigated. Starting from depth $D = 2$, meaning a single link between centroid and leaves, the plateaux curve is obtained running the algorithm for values of lambda included into the range $\lambda \in [0.1, 0.8]$ and spaced 0.01. Figure 4 shows the passage from nearly no clusters, for λ close to 0, to a single cluster for λ large. The blue line shows the mode of number of clusters, while the red one is the average number of clusters with its standard deviation. An example of clustering obtained for lambdas falling in the specific plateaux range is shown; blue bullets represents healthy individuals, while red ones are individuals affected by the pathology.

The 5-clusters plauteaux, letter a) in figure 4, presents three healthy individuals clusters and two cleft lip clusters. Observing labels following bullets (not shown in the figure), one can analyze deeper the structure of the two red clusters and it is possible to appreciate how they are divided according to the presence of right- and left-sided cleft lip (see also figure 7). Proceeding to larger values of lambda, the range highlighted with letter b) indicates a region in which two of the healthy individuals clusters merge together and then merge again with both the cleft-lip clusters (letter c)). From this analysis it turns out how 2-MST unsupervised clustering is not able to unveil the inner structure of the proposed data and, trying to impose a spherical geometry, it returns more than one sub-population for the healthy individual class.

The same procedure is repeated moving to depth $D = 3$. Again the parameter λ is spanned from the minimum to the maximum of the dissimilarities, looking for the largest stability region of the algorithm. With higher clustering depths, it is found how the decay toward a single cluster plateaux is faster with respect to the spherical clustering. Thus, in order to build a reliable plateaux curve, a finer grating for lambda values is required, especially in the transition region. Moreover, at this depth it is quite common to encounter outliers, i.e. single nodes that are not assigned to any class. To this purpose, a green line is plotted as well, showing the mode of number of clusters composed by at least two nodes, in order to unveil the number of outliers present in the clustering.

Figure 5 shows the plateaux curve for depth $D = 3$. In such a case, two plateaux with no outliers are identified in the transition region between none and one cluster. Here lambda spans with a 0.002 step in order to track in detail the plateaux behaviour, while in the other regions of the curve a 0.005 step is kept. The most important parameter region, i.e. that with three clusters and indicated with letter b) in figure 5, discriminates well healthy individuals from those manifesting cleft lip. In particular, left- and right-sided cleft lips are clustered in two separated classes, while healthy individuals create a unique cluster. Then, the algorithm succeeds in identifying the investigated structure of the proposed data.

In order to understand the robustness of the clustering against the stochasticity of the algorithm, we compute the probability for any node of being assigned to its own class. Setting the value of lambda inside the plateaux, $\lambda = 0.18$, the algorithm is run fifty times, building a statistics of the

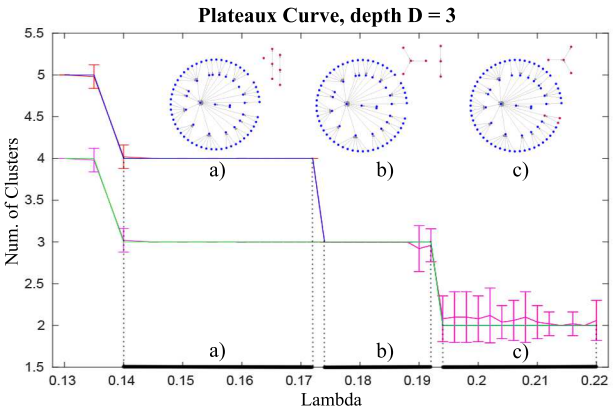


Figure 5: Number of clusters versus λ, for $D = 3$. Plateaux curve. The green solid line represents the obtained plateaux curve of the resulting clusters with at least 2 nodes. A single plateau, i.e. plateau b), is stable in converging toward a solution with no outliers and reports the discrimination of individuals with a pathology (see figure 7).

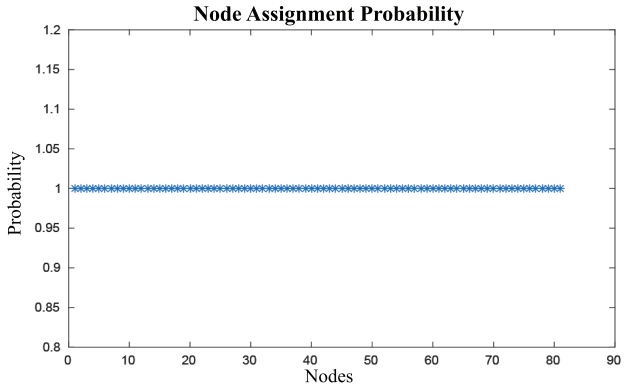


Figure 6: Assignment Probability Plot. The probability of a node to be assigned to its most frequent cluster throughout 50 runs of the clustering algorithm is plotted. No value below 1 is found.

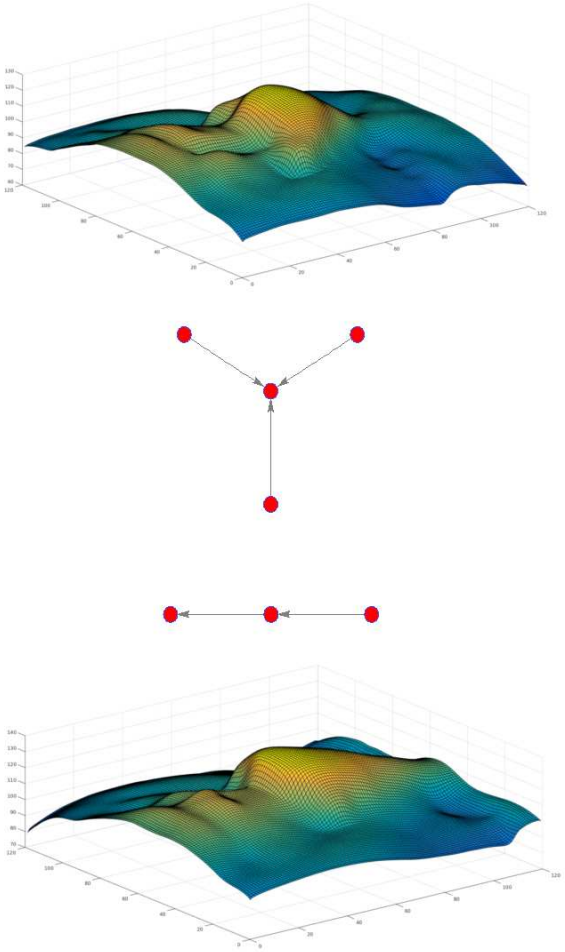


Figure 7: Resulting Clusters For $D = 3$. The 4-nodes cluster includes all the right-sided cleft lip affected individuals; an example of such individual surface is reported at the top of the figure. On the other hand, the 3-nodes cluster includes left-sided cleft lip affected individuals. An example of individual surface belonging to this class is reported at the bottom of the figure.

class assignment for any node. Plotting the node assignment probability, as shown in figure 6, it can be observed how all individuals show a probability of being member of their most frequent class equal to 1. Moreover, comparing each clustering returned in subsequent runs with the first one obtained, we can compute the overlap among clusterings, finding that the algorithm converges always to the same classifications, in terms of composition of the single clustering. In conclusion, such a clustering overlap and assignment probability allow us to state with high confidence the robustness of the classification returned.

Higher depths trees are also investigated, but results are not reported here, as they do not improve sensibly those obtained for $D = 3$. This means that, once the spherical constraint is relaxed, data do not show a longer range dependency and a three nodes correlation is able to catch the inner structure of clusters.

4 Conclusions

This work proposes an innovative automatic methodology for the diagnosis of cleft lip. The process is designed for the detection of diseases in the prenatal phase, thus its feasibility is tested upon the limitedly available ultrasound scans data and then deeply investigated for a large dataset of adult individuals. Bosphorus database is chosen for the testing, which seven cleft lip-affected individuals are added to, by artificially imposing the defect.

The algorithm maps each face with twelve differential geometry descriptors plus first and second derivatives with respect to x and y directions. It allows to determine facial landmarks, that enable face comparison. In this work, only the *pronasal* and the *labrum superior* landmarks are investigated. They are identified automatically by imposing thresholds on values expressed by a subset of geometrical descriptors. In particular, the *labrum superior* specifies the region of interest for the actual diagnosis.

In the second part of the algorithm, the geometrical descriptors expressed in the *labrum superior*'s neighbourhood are used to transform each face into a vector and create the input matrix for the unsupervised clustering algorithm. Any entrance of such built vector is used to perform the comparison between couples of individuals and compute their dissimilarity in terms of Spearman's correlation rank distance. In such a way, a squared symmetric matrix is computed and provided as input for the clustering itself.

Eventually, D-MST clustering algorithm allows to investigate regions of convergence stability to a certain number of clusters, the so called plateaux, imposing the maximum depth, i.e. the inner dependencies structure, of any cluster detected. It correctly separates left-sided and right-sided cleft lips, thus showing accurate diagnosis results.

This algorithm also opens the route for the definition of what is called *normotype*. The normotype can be considered as the representative face of a class of individuals,

collecting all the principal features distinguishing an individual as member of a particular category. The present algorithm is able to collect all healthy individuals in a single cluster, starting from features expressed in the lip region, in comparison to those manifesting a cleft lip. This allows a formalization of the normotype features. On the other hand, once the cleft lip population will account for a sufficient number of members, the labio-schisis normotype would be defined as well.

Other syndromes, like the Fetal Alchol Syndrome (FAS) or the palato-schisis syndrome, will also be investigated and geometrically formalized to be embedded in this algorithm, which could be a multi-syndrome diagnosing tool.

References

[1] T. S. Douglas, T. E. M. Mutsvangwa, "A Review of Facial Image Analysis for Delineation of the Facial Phenotype Associated With Fetal Alcohol Syndrome," *Am J Med Genet Part A*, vol. 152, no. 2, pp. 528536, 2010.

[2] S. Campbell, C. Lees, G. Moscoso, P. Hall, "Ultrasound antenatal diagnosis of cleft palate by a new technique: the 3D reverse face view," *Ultrasound Obstet Gynecol*, vol. 25, no. 1, pp. 1218, 2005.

[3] I. Aras, S. Olmez, S. Dogan, "Comparative Evaluation of Nasopharyngeal Airways of Unilateral Cleft Lip and Palate Patients Using Three-Dimensional and Two-Dimensional Methods," *The Cleft Palate-Craniofacial Journal*, vol. 49, no. 6, pp. e75e81, 2012.

[4] L. D. Platt, G. R. DeVore, D. H. Pretorius, "Improving Cleft Palate/Cleft Lip Antenatal Diagnosis by 3-Dimensional Sonography," *J Ultrasound Med*, vol. 25, no. 11, pp. 14231430, 2006.

[5] G. Tonni, M. Lituania, "OmniView Algorithm, A Novel 3-Dimensional Sonographic Technique in the Study of the Fetal Hard and Soft Palates," *J Ultrasound Med*, vol. 31, no. 2, pp. 313318, 2012.

[6] P. A. Mossey, J. Little, R. G. Munger, M. J. Dixon, W. C. Shaw, "Cleft lip and palate," *The Lancet*, vol. 374, no. 9703, pp. 1773-1785, 2009.

[7] L. Bonacina, D. Conti, A. Froio, F. Marcolin, E. Vezzetti, "Automatic 3D Fetus Face Model Extraction From Ultrasonography Through Histogram Processing", submitted to *IEEE Trans Biomed Eng*.

[8] A. K. Jain, M. N. Murty, P. J. Flinn, "Data Clustering: A Review", *ACM Computing Surveys*, vol. 31, no. 3, 1999.

[9] B. J. Frey, D. Dueck, "Clustering by Passing Messages Between Data Points," *Science*, vol. 315, no. 972, 2007.

[10] J.H. Jr. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236-244, 1963.

[11] M. Bailly-Bechet, S. Bradde, A. Braunstein, A. Flaxman, L. Foini, R. Zecchina, "Clustering with shallows trees," *Journal of Statistical Mechanics: Theory and Experiments*, vol. 2009, doi:10.1088/1742-5468/2009/12/P12010, 2009.

[12] M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J.-M. Franois, and R. Zecchina. "Finding undetected protein associations in cell signaling by belief propagation," *Proceedings of the National Academy of Sciences*. 2010; 108(2):882-7, doi:10.1073/pnas.1004751108

[13] The Bosphorus Database Official Web Site: *http://bosphorus.ee.boun.edu.tr/default.aspx*

[14] E. Vezzetti, F. Marcolin, "Geometrical descriptors for human face morphological analysis and recognition," *Robotics and Autonomous Systems*, vol. 60, no. 6, pp. 928-939, 2012.

[15] J. J. Koenderick, A. J. van Doorn, "Surface shape and curvature scales", *Image and Vision Computing*, vol. 10, no. 8, pp. 557-564, 1992.

[16] E. Vezzetti, D. Speranza, F. Marcolin, G. Fracastoro, G. Buscicchio, "Exploiting 3D Ultrasound for Fetal Diagnostic Purpose Through Facial Landmarking," *Image Anal Stereol*, vol. 33, no. 3, pp. 167-188, 2014.

[17] E. Vezzetti, D. Speranza, F. Marcolin, G. Fracastoro, "Diagnosing cleft lip pathology in 3d ultrasound: A landmarking-based approach," *Image Anal Stereol*, vol. 35, no. 1, pp. 53-65, 2015.

[18] S. Moos, F. Marcolin, S. Tornincasa, E. Vezzetti, M. G. Violante, G. Fracastoro, D. Speranza, F. Padula, "Cleft lip pathology diagnosis and foetal landmark extraction via 3D geometrical analysis," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 11, no. 1, pp. 1-18, 2017.

[19] E. Vezzetti, F. Marcolin, "Geometry-based 3D face morphology analysis: soft-tissue landmark formalization," *Multimed Tools Appl*, vol. 68, no. 3, pp. 895-929, 2012.

[20] E. Vezzetti, F. Marcolin, "3D human face description: landmarks measures and geometrical features," *Image and Vision Computing*, vol. 30, no. 10, pp. 698-712, 2012.

[21] E. Vezzetti, S. Moos, F. Marcolin, V. Stola, "A pose-independent method for 3D face landmark formalization," *Computer Methods and Programs in Biomedicine*, vol. 198, no. 3, pp. 1078-1096, 2012.

[22] E. Vezzetti, F. Marcolin, V. Stola, "3D Human Face Soft Tissues Landmarking Method: An Advanced Approach," *Computers in Industry*, vol. 64, no. 9, pp. 1326-1354, 2013.

[23] E. Vezzetti, F. Marcolin, "3D Landmarking in Multiexpression Face Analysis: A Preliminary Study on Eyebrows and Mouth," *Aesthetic Plastic Surgery*, vol. 38, pp. 796-811, 2014.

[24] R. Xu, C. Donald, *Clustering*, Wiley-IEEE Press, November 2008.

[25] M. Bayati, A. Braunstein, R. Zecchina, "A rigorous analysis of the cavity equations for the minimum spanning tree," *Journal of mathematical physics*, vol. 49, no. 12, 2008.

[26] M. Bayati *et al.*, "Statistical mechanics of Steiner trees," *Physical Review Letters*, vol 101, no. 3, 2008.

Computational Intelligence Algorithms for the Development of an Artificial Sport Trainer

Iztok Fister Jr.
University of Maribor, Faculty of Electrical Engineering and Computer Science, Koroška 34, 2000 Maribor
E-mail: iztok.fister1@um.si

Thesis summary

Keywords: computational intelligence, nature-inspired algorithms, sport training

Received: October 12, 2017

This paper presents a short summary of doctoral thesis that proposes the use of computational intelligence algorithms for the development of an Artificial Sport Trainer.

Povzetek: Članek predstavlja kratko vsebino doktorske disertacije, ki predlaga uporabo algoritmov računske inteligence za razvoj umetnega športnega trenerja.

1 Introduction

Planning the proper sport training sessions for athletes is a very hard problem for sports trainers. With the rising computational power on the one hand, and emerging data warehouses on the other, new algorithms for discovering knowledge from these data have emerged. An overview of literature in this domain showed that there is still a lack of algorithms for knowledge enrichment from data that are based explicitly on computational intelligence [1]. Interestingly, there were some solutions that applied artificial neural networks for controlling the sports training sessions (e.g. [5, 6]) and even fuzzy logic [7]. However, almost none of the studies suggested the use of population-based nature-inspired algorithms [4] (e.g., evolutionary algorithms, or swarm intelligence algorithms) for these tasks. Contrary, the thesis [3] proposes a concept of an intelligent system called **Artificial Sport Trainer** (AST) for the training purposes of athletes. AST is based on stochastic population-based nature-inspired algorithms, designed to cover all phases of the sports training, which are also in the domain of a real sport trainer: planning, realization, control and evaluation.

The thesis is divided into two parts. The first, theoretical part, presents the fundamentals of computational intelligence, basics of sports training and the architecture of the AST. The second, experimental part, presents two applications of the AST. The former is devoted to planning sports training sessions based on existing sports activities, e.g. comprehensive performance study of six different stochastic, nature-inspired population-based algorithms (Bat Algorithm (BA), Differential Evolution (DE), Firefly Algorithm (FA), Hybrid Bat Algorithm (HBA), self-adaptive Differential Evolution (jDE) and Particle Swarm Optimization (PSO)). These algorithms were tested on three real datasets, i.e. professional cyclist, amateur cyclist and semi-professional runner. The latter proposes a solution for Association Rule Mining (ARM) based on BA (the so-called BatMiner), which is applied to real datasets for finding cyclist's characteristics during the sports training.

2 Artificial sport trainer

The main architecture of the AST [2] covers the following phases of the sport training:

- Planning: the most important phase of the sport training, that consists of:
  - long-term planning (so-called strategy) and
  - short-term planning (so-called tactics).
- Realization: this phase captures the realization of the sports training session.
- Control: realization of the sports sessions is typically controlled by wearable devices, such as sports watches or smart-phones.
- Evaluation: after the conducted training plan, the expected form or abilities of an athlete are evaluated.

3 Experiments and results

In order to confirm that the AST can be used in practice, we have conducted a comprehensive experimental work that includes mentioned six different algorithms (i.e., BA, DE, FA, HBA, jDE, PSO) on three real datasets obtained by different kinds of athletes (i.e., professional, semi-professional and amateur) in two sports (i.e. cycling, running). Additionally, we have also studied the influence of clusters that was obtained by k-means clustering. We have used the following numbers of clusters: 5, 8, 10, 12,

15, and 18. The second application was tested on a real cyclist's dataset and was compared to the Hybrid Binary Cuckoo Search for ARM. Resulting plans of the first application have then been compared to the plans, created by a real sport trainer. Comparison showed that the AST can be used for planning sport trainings sessions, according to the TRIMP indicator with confidence of 0.1. The results of the second application showed that a BatMiner is an appropriate algorithm for finding characteristics of athletes during the sports training.

4 Conclusion

Main findings of the thesis [3] are: (1) A new research area is proposed, i.e., use of computational intelligence algorithms in the sport area, (2) The concept of an **Artificial Sport Trainer** encompasses various algorithms of computational intelligence in sport, (3) New population-based nature-inspired algorithms for planning sport training sessions are developed and validated on the real data obtained by two cyclists and one runner, (4) An easy metric for comparing AST's and real trainer's session plans is proposed and (5) The BatMiner algorithm for mining characteristics of athletes during the sports training sessions is built.

References

[1] Andries P Engelbrecht. *Computational intelligence: an introduction*. John Wiley & Sons, 2007.

[2] Iztok Fister, Karin Ljubič, Ponnuthurai Nagaratnam Suganthan, and Matjaž Perc. Computational intelligence in sports: challenges and opportunities within a new research domain. *Applied Mathematics and Computation*, 262:178–186, 2015.

[3] Iztok Fister Jr. Algoritmi računske inteligence za razvoj umetnega športnega trenerja. Doctoral thesis, University of Maribor, Slovenia, 2017.

[4] Iztok Fister Jr, Xin-She Yang, Iztok Fister, Janez Brest, and Dušan Fister. A brief review of nature-inspired algorithms for optimization. *arXiv preprint arXiv:1307.4186*, 2013.

[5] Hristo Novatchkov and Arnold Baca. Machine learning methods for the automatic evaluation of exercises on sensor-equipped weight training machines. *Procedia Engineering*, 34:562–567, 2012.

[6] Hristo Novatchkov and Arnold Baca. Artificial intelligence in sports on the example of weight training. *Journal of sports science & medicine*, 12(1):27, 2013.

[7] Hristo Novatchkov and Arnold Baca. Fuzzy logic in sports: a review and an illustrative case study in the field of strength training. *International Journal of Computer Applications*, 71(6), 2013.

CALL FOR PAPERS

Information Society 2018

21th International Multiconference

8–12 October 2018, Ljubljana, Slovenia

IS 2018

http://is.ijs.si

The concepts of information society, information era, infosphere and infostress have by now been widely accepted including futuristic ideas about emerging superintelligence. But what does it really mean for the society, science, technology, education, governments, our lives? The Information Society multiconference deals with information technologies, which are of major importance for the development of Europe and the world.

Information Society 2018 will serve as a forum for the world-wide and national community to explore further research directions, business opportunities and governmental policies. For these reasons we host a scientific meeting in the form of a multiconference, which will consist of several independent conferences with themes essential for the development of the information society. The main objective is the exchange of ideas and developing visions for the future of information society. IS 2018 is a high-quality multidisciplinary conference covering major recent scientific achievements.

From the best papers, presented at the Information society conference, a special issue will be assembled in the Informatica journal.

Informatica (http://www.informatica.si) is an international journal with its base in Europe. It publishes peer-reviewed papers from all areas of computer and information science: mostly scientific and technical, but also commercial and industrial. Informatica has been published continuously since 1976 by Slovenian Society Informatika. It is indexed in several databases including Thomson Reuters' Emerging Sources Citation Index.

For submission dates, please check each conference details on https://is.ijs.si/. However, September 1, 2018 is an orientation deadline for an average IS2018 conference.

CONTENTS OF *Informatica* Volume 41 (2017) pp. 1–523

Papers

Editorials

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of **Slove**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please register as an author and submit a manuscript at: http://www.informatica.si. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

SUBSCRIPTION

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than twentythree years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica web edition is free of charge and accessible at http://www.informatica.si.
Informatica print edition is free of charge for major scientific, educational and governmental institutions. Others should subscribe.

Informatica WWW:

http://www.informatica.si/

Referees from 2008 on:

A. Abraham, S. Abraham, R. Accornero, A. Adhikari, R. Ahmad, G. Alvarez, N. Anciaux, R. Arora, I. Awan, J. Azimi, C. Badica, Z. Balogh, S. Banerjee, G. Barbier, A. Baruzzo, B. Batagelj, T. Beaubouef, N. Beaulieu, M. ter Beek, P. Bellavista, K. Bilal, S. Bishop, J. Bodlaj, M. Bohanec, D. Bolme, Z. Bonikowski, B. Bokovi, M. Botta, P. Brazdil, J. Brest, J. Brichau, A. Brodnik, D. Brown, I. Bruha, M. Bruynooghe, W. Buntine, D.D. Burdescu, J. Buys, X. Cai, Y. Cai, J.C. Cano, T. Cao, J.-V. Capella-Hernndez, N. Carver, M. Cavazza, R. Ceylan, A. Chebotko, I. Chekalov, J. Chen, L.-M. Cheng, G. Chiola, Y.-C. Chiou, I. Chorbev, S.R. Choudhary, S.S.M. Chow, K.R. Chowdhury, V. Christlein, W. Chu, L. Chung, M. Ciglari, J.-N. Colin, V. Cortellessa, J. Cui, P. Cui, Z. Cui, D. Cutting, A. Cuzzocrea, V. Cvjetkovic, J. Cypryjanski, L. ehovin, D. erepnalkoski, I. osi, G. Daniele, G. Danoy, M. Dash, S. Datt, A. Datta, M.-Y. Day, F. Debili, C.J. Debono, J. Dedi, P. Degano, A. Dekdouk, H. Demirel, B. Demoen, S. Dendamrongvit, T. Deng, A. Derezinska, J. Dezert, G. Dias, I. Dimitrovski, S. Dobriek, Q. Dou, J. Doumen, E. Dovgan, B. Dragovich, D. Drajic, O. Drbohlav, M. Drole, J. Dujmovi, O. Ebers, J. Eder, S. Elaluf-Calderwood, E. Engstrm, U. riza Erturk, A. Farago, C. Fei, L. Feng, Y.X. Feng, B. Filipi, I. Fister, I. Fister Jr., D. Fier, A. Flores, V.A. Fomichov, S. Forli, A. Freitas, J. Fridrich, S. Friedman, C. Fu, X. Fu, T. Fujimoto, G. Fung, S. Gabrielli, D. Galindo, A. Gambarara, M. Gams, M. Ganzha, J. Garbajosa, R. Gennari, G. Georgeson, N. Gligori, S. Goel, G.H. Gonnet, D.S. Goodsell, S. Gordillo, J. Gore, M. Grar, M. Grgurovi, D. Grosse, Z.-H. Guan, D. Gubiani, M. Guid, C. Guo, B. Gupta, M. Gusev, M. Hahsler, Z. Haiping, A. Hameed, C. Hamzaebi, Q.-L. Han, H. Hanping, T. Hrder, J.N. Hatzopoulos, S. Hazelhurst, K. Hempstalk, J.M.G. Hidalgo, J. Hodgson, M. Holbl, M.P. Hong, G. Howells, M. Hu, J. Hyvrinen, D. Ienco, B. Ionescu, R. Irfan, N. Jaisankar, D. Jakobovi, K. Jassem, I. Jawhar, Y. Jia, T. Jin, I. Jureta, . Jurii, S. K, S. Kalajdziski, Y. Kalantidis, B. Kalua, D. Kanellopoulos, R. Kapoor, D. Karapetyan, A. Kassler, D.S. Katz, A. Kaveh, S.U. Khan, M. Khattak, V. Khomenko, E.S. Khorasani, I. Kitanovski, D. Kocev, J. Kocijan, J. Kollr, A. Kontostathis, P. Koroec, A. Koschmider, D. Koir, J. Kova, A. Krajnc, M. Krevs, J. Krogstie, P. Krsek, M. Kubat, M. Kukar, A. Kulis, A.P.S. Kumar, H. Kwanicka, W.K. Lai, C.-S. Laih, K.-Y. Lam, N. Landwehr, J. Lanir, A. Lavrov, M. Layouni, G. Leban, A. Lee, Y.-C. Lee, U. Legat, A. Leonardis, G. Li, G.-Z. Li, J. Li, X. Li, X. Li, Y. Li, Y. Li, S. Lian, L. Liao, C. Lim, J.-C. Lin, H. Liu, J. Liu, P. Liu, X. Liu, X. Liu, F. Logist, S. Loskovska, H. Lu, Z. Lu, X. Luo, M. Lutrek, I.V. Lyustig, S.A. Madani, M. Mahoney, S.U.R. Malik, Y. Marinakis, D. Marini, J. Marques-Silva, A. Martin, D. Marwede, M. Matijaevi, T. Matsui, L. McMillan, A. McPherson, A. McPherson, Z. Meng, M.C. Mihaescu, V. Milea, N. Min-Allah, E. Minisci, V. Mii, A.-H. Mogos, P. Mohapatra, D.D. Monica, A. Montanari, A. Moroni, J. Mosegaard, M. Mokon, L. de M. Mourelle, H. Moustafa, M. Moina, M. Mrak, Y. Mu, J. Mula, D. Nagamalai, M. Di Natale, A. Navarra, P. Navrat, N. Nedjah, R. Nejabati, W. Ng, Z. Ni, E.S. Nielsen, O. Nouali, F. Novak, B. Novikov, P. Nurmi, D. Obrul, B. Oliboni, X. Pan, M. Panur, W. Pang, G. Papa, M. Paprzycki, M. Parali, B.-K. Park, P. Patel, T.B. Pedersen, Z. Peng, R.G. Pensa, J. Per, D. Petcu, B. Petelin, M. Petkovek, D. Pevec, M. Piulin, R. Piltaver, E. Pirogova, V. Podpean, M. Polo, V. Pomponiu, E. Popescu, D. Poshyvanyk, B. Potonik, R.J. Povinelli, S.R.M. Prasanna, K. Pripui, G. Puppis, H. Qian, Y. Qian, L. Qiao, C. Qin, J. Que, J.-J. Quisquater, C. Rafe, S. Rahimi, V. Rajkovi, D. Rakovi, J. Ramaekers, J. Ramon, R. Ravnik, Y. Reddy, W. Reimche, H. Rezankova, D. Rispoli, B. Ristevski, B. Robi, J.A. Rodriguez-Aguilar, P. Rohatgi, W. Rossak, I. Roanc, J. Rupnik, S.B. Sadkhan, K. Saeed, M. Saeki, K.S.M. Sahari, C. Sakharwade, E. Sakkopoulos, P. Sala, M.H. Samadzadeh, J.S. Sandhu, P. Scaglioso, V. Schau, W. Schempp, J. Seberry, A. Senanayake, M. Senobari, T.C. Seong, S. Shamala, c. shi, Z. Shi, L. Shiguo, N. Shilov, Z.-E.H. Slimane, F. Smith, H. Sneed, P. Sokolowski, T. Song, A. Soppera, A. Sorniotti, M. Stajdohar, L. Stanescu, D. Strnad, X. Sun, L. ajn, R. enkek, M.R. ikonja, J. ilc, I. krjanc, T. tajner, B. ter, V. truc, H. Takizawa, C. Talcott, N. Tomasev, D. Torkar, S. Torrente, M. Trampu, C. Tranoris, K. Trojacanec, M. Tschierschke, F. De Turck, J. Twycross, N. Tziritas, W. Vanhoof, P. Vateekul, L.A. Vese, A. Visconti, B. Vlaovi, V. Vojisavljevi, M. Vozalis, P. Vraar, V. Vrani, C.-H. Wang, H. Wang, H. Wang, H. Wang, S. Wang, X.-F. Wang, X. Wang, Y. Wang, A. Wasilewska, S. Wenzel, V. Wickramasinghe, J. Wong, S. Wrobel, K. Wrona, B. Wu, L. Xiang, Y. Xiang, D. Xiao, F. Xie, L. Xie, Z. Xing, H. Yang, X. Yang, N.Y. Yen, C. Yong-Sheng, J.J. You, G. Yu, X. Zabulis, A. Zainal, A. Zamuda, M. Zand, Z. Zhang, Z. Zhao, D. Zheng, J. Zheng, X. Zheng, Z.-H. Zhou, F. Zhuang, A. Zimmermann, M.J. Zuo, B. Zupan, M. Zuqiang, B. alik, J. ika,

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Litostrojska cesta 54, 1000 Ljubljana, Slovenia.
The subscription rate for 2017 (Volume 41) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.
Printing: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Slovene Society for Pattern Recognition (Simon Dobrišek)
Slovenian Artificial Intelligence Society (Mitja Luštrek)
Cognitive Science Society (Olga Markič)
Slovenian Society of Mathematicians, Physicists and Astronomers (Marej Brešar)
Automatic Control Society of Slovenia (Nenad Muškinja)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Stane Pejovnik)
ACM Slovenia (Matjaž Gams)

Informatica is financially supported by the Slovenian research agency from the Call for co-financing of scientific periodical publications.

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

Volume 41 Number 4 December 2017 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Informatica **41** (2017) Number 4, pp. 383–523