CONCEPTUALIZING AI RISK

Martin Ciupa[1] and Keith Abney[2]

[1]CTO calvIO Inc., Webster, New York, USA
[2]Senior Lecturer, CalPoly, California, USA

ABSTRACT

AI advances represent a great technological opportunity, but also possible perils. This paper undertakes an ethical and systematic evaluation of those risks in a pragmatic analytical form of questions, which we term 'Conceptual AI Risk analysis'. We then look at a topical case example in an actual industrial setting and apply that methodology in outline. The case involves Deep Learning Black-Boxes and their risk issues in an environment that requires compliance with legal rules and industry best practices. We examine a technological means to attempt to solve the Black-box problem for this case, referred to as "Really Useful Machine Learning" (RUMLSM). DARPA has identified such cases as being the "Third Wave of AI." Conclusions to its efficacy are drawn.

KEYWORDS

AI Risk, Deep Neural Network, Black-box Problem, Really Useful Machine Learning, RUMLSM, DARPA Third Wave of AI.

1. INTRODUCTION

A common worry about AI is that it poses an unacceptable risk to humanity (or individual humans) in some way. An extensive literature has begun to emerge about various aspects of Artificial Intelligence (AI) risk, much of it focused on existential risk from Artificial Generic Intelligence (AGI). But AI poses other risks, from how driverless cars solve the 'trolley problem', to whether autonomous military robots attack only legitimate targets, to trust in the safety of AI/Robotics in industrial and commercial settings. More generally, the discussion of risks from AI has paid insufficient attention to the nature of risk itself, as well as how decisions about the acceptability of the risks of AI compare to worries about convergent technologies. For example, global debates about autonomous weapons have focused predominantly on robotics, but AI also can be weaponized. For instance, in robotics serious concern exists over a possible lack of "meaningful human control" [1]. Missing is a similar concern for autonomous AI-controlled cyber attacks that would lack the very same control [2]. The Vice Chairman of the Joint Chiefs of Staff understands, saying, "In the [Defense] Department, we build machines and we test them until they break. You can't do that with an artificial intelligence, deep learning piece of software. We're going to have to figure out how to get the software to tell us what it's learned" [3]. Significant aspects of AI risk can thus be characterized as a "black-box" problem.

Natarajan Meghanathan et al. (Eds) : CCSEA, NCOM, AIFU, DKMP, EMSA, SIPRO, SEA - 2018
pp. 73– 81, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80307

So, how best to understand the risks of AI, judge them (un)acceptable, and then apply our insights on risk to determine what policies to pursue?

2. DEFINING RISK, AND HOW TO THINK ABOUT IT

So, AI poses many different types of risk – but what exactly is risk? Andrew Maynard [4] suggests that we start with the idea of "value." If innovation is defined as creating value that someone is willing to pay for, then he suggests risk as a *threat to value*, and not just in the ways value is usually thought of when assessing risk, such as health, the environment or financial gain/loss, but also well-being, environmental sustainability, deeply held beliefs, even a sense of cultural or personal identity. Risk is thus a potential, but not certain, harm.

Extending a schema based on previous work [5], the major factors in determining 'acceptable risk' in AI will include (but are not limited to):

2.1 Acceptable-Risk Factor: Consent

Consent: Is the risk voluntarily endured, or not? For instance, secondhand smoke is generally more objectionable than firsthand, because the passive smoker did not consent to the risk, even if the objective risk is smaller. Will those who are at risk from AI reasonably give consent? When would it be appropriate to deploy or use AI without the meaningful consent of those affected? Would non-voluntariness (in which the affected party is unaware of the risk/ cannot consent) be morally different from involuntariness (in which the affected party is aware of the risk and does not consent)? [6]

2.2 Acceptable-Risk Factor: Informed Consent

Even if AIs only have a 'slave morality' in which they always follow orders [7], and citizens consent to their use (through, say, political means), that still leaves unanswered whether the risk (of malfunction, unintended consequences, or other error) to *unintended* parties is morally permissible. After all, even if widespread consent is in some sense possible, it is completely unrealistic to believe that all humans affected by AI could give *informed* consent to their use. So, does the morality of consent require adequate knowledge of what is being consented to?

Informed consent: Are those who undergo the risk voluntarily fully aware of the true nature of the risk? Or would such knowledge undermine their efficacy in fulfilling their (risky) roles? Or are there other reasons for preferring ignorance? Thus, will all those at risk from AI know that they are at risk? If not, do those who know have an obligation to inform others of the risks? What about foreseeable but unknown risks—how should they (the 'known unknowns') be handled? Could informing people that they are at risk ever be unethical, even akin to terrorism?

2.3 Acceptable-Risk Factor: The Affected Population

Even if consent or informed consent do not appear to be morally required with respect to some AI, we may continue to focus on the affected population as another factor in determining acceptable risk:

Affected population: Who is at risk—is it merely groups that are particularly susceptible or innocent, or those who broadly understand that their role is risky, even if they do not know the particulars of the risk? For example, in military operations civilians and other noncombatants are usually seen as not morally required to endure the same sorts of risks as military personnel, even (or especially) when the risk is involuntary or non-voluntary.

2.4 Acceptable-Risk Factor: Step risk versus State risk

A state risk is the risk of being in a certain state, and the total amount of risk to the system is a direct function of the time spent in the state. Thus, state risk is time-dependent; total risk depends (usually linearly) on the time spent in the state. So, for us living on the surface of the Earth, the risk of death by asteroid strike is a state risk (it increases the longer we're here).

Step risk, on the other hand, is a discrete risk of taking the next step in some series or undergoing some transition; once the transition is complete, the risk vanishes. In general, step risk is not time-dependent, so the amount of time spent on step matters little (or not at all). [8] Crossing a minefield is usually a step risk – the risk is the same whether you cross it in 1 minute or 10 minutes. The development of AGI poses a clear step risk; but, if there is a 'fast takeoff,' the state risk may be negligible.

2.5 Acceptable-Risk Factors: Seriousness and Probability

We thereby come to the two most basic facets of risk assessment, seriousness, and probability: how bad would the harm be, and how likely is it to happen?

Seriousness: A risk of death or serious physical (or psychological) harm is understandably seen differently than the risk of a scratch or a temporary power failure or slight monetary costs. But the attempt to make serious risks nonexistent may turn out to be prohibitively expensive. What (if any) serious risks from AIs are acceptable—and to whom: users, nonusers, the environment, or the AI itself?

Probability: This is sometimes conflated with seriousness but is intellectually quite distinct. The seriousness of the risk of a 10-km asteroid hitting Earth is quite high (possible human extinction), but the probability is reassuringly low (though not zero, as perhaps the dinosaurs discovered). What is the probability of harm from AIs? How much certainty can we have in estimating this probability? What probability of serious harm is acceptable? What probability of moderate harm is acceptable? What probability of mild harm is acceptable?

2.6 Acceptable-Risk Factors: Who Determines Acceptable Risk?

In various other social contexts, all of the following have been defended as proper methods for determining that a risk is unacceptable [9]:

Good faith subjective standard: It is up to each individual as to whether an unacceptable risk exists. That would involve questions such as the following: Can the designers or users of AI be trusted to make wise choices about (un)acceptable risk? The idiosyncrasies of human risk aversion may make this standard impossible to defend, as well as the problem of involuntary/ non-voluntary risk borne by nonusers.

The reasonable-person standard: An unacceptable risk is simply what a fair, informed member of a relevant community believes to be an unacceptable risk. Can we substitute a professional code or some other basis for what a 'reasonable person' would think for the difficult-to-foresee vagaries of conditions in the rapidly emerging AI field, and the subjective judgment of its practitioners and users? Or what kind of judgment would we expect an autonomous AI to have—would we trust it to accurately determine and act upon the assessed risk? If not, then can AI never be deployed without teleoperators—like military robots, should we always demand a human in the loop? But even a 'kill switch' that enabled autonomous operation until a human doing remote surveillance determined something had gone wrong would still leave unsolved the first-generation problem.

Objective standard: An unacceptable risk requires evidence and/or expert testimony as to the reality of (and unacceptability of) the risk. But there remains the first-generation problem: how do we understand that something is an unacceptable risk unless some first generation has already endured and suffered from it? How else could we obtain convincing objective evidence?

2.7 Acceptable-Risk Factors: The Wild Card: Existential Risk?

Plausibly, a requirement for extensive, variegated, realistic, and exhaustive pre-deployment testing of AIs in virtual environments before they are used in actual human interactions could render many AI risks acceptable under the previous criteria. But one AI risk may remain unacceptable even with the most rigorous pre-deployment testing. An existential risk refers to a risk that, should it come to pass, would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential. Existential disasters would end human civilization for all time to come. For utilitarians, existential risks are terribly important: doing what we can to mitigate even a small chance that humanity comes to an end may well be worth almost any cost. And for deontologists, the idea that 'one always has a moral obligation never to allow the extinction of all creatures capable of moral obligation' is at least a plausible *prima facie* (and perhaps absolute) duty; such a survival principle appears required for any viable ethics [10]. If there is even a tiny risk that developing AGI would pose an existential risk, this 'Extinction Principle' may well imply that we have a duty to stop it.

3. SPECIFIC CASE STUDY, POSSIBLE SOLUTION, AND RISK ANALYSIS

3.1 Specific Case Study

As noted above, a well-known problem of AI risk is applying Deep Learning Neural Networks, specifically the black-box problem. This occurs when the Neural Network has a large set of training data; the consequential connectivity arrangements are such that for a given input, an output is provided to a reinforced goal of the system – typically prediction, decision or identification. These bottom-up systems can learn well, especially in tight domains, and in them can surpass human performance – but, without any means of validation, or clear compliance to regulations. This renders the AI potentially untrustworthy, a serious deployment risk that may be deemed unacceptable. A human can usually provide a top-down rationalization of their behavior, responding to "why did you do that" questions. But a Deep Learning system cannot easily answer such queries. It is not rule-based and cannot easily track its "reasoning." [11]

We propose a possible solution. Our case study is a Deep Learning system applied to the Path Planning of a Robot Arm, in which vials of biohazardous materials are to be moved from point A to point B in an optimum path. This path is constrained by parameters such as speed, power-usage, minimization of actuator acceleration and deceleration (that causes wear of the actuators) and collision avoidance. See Fig 1. Keep in mind that cost of production, as well as quality/safety, are value factors to be balanced in this manufacturing example. And the use of AI Robots in this case example is a very real-world example of potential benefit and ethical concerns.

A key part of the solution proposed is an "Extractor" which will build a rationalization of the Deep Learning System into a Rule-Based Decision tree that can be validated against risk analysis/compliance needs, i.e., answering questions related to risks. This is depicted in Fig 2.

3.2 Possible Solution: Extraction of Heuristics from Deep Learning Neural Net

The means by which Expert Heuristics are extracted from the Deep Learning Neural Networks has been studied by other teams [12], [13] and [14]. The specific means by which we propose to do so in RUML^{SM} is an innovative patent pending process [15], [16] and [17]. Expert Heuristic/Rule extraction can be defined as "…given a trained neural network and the data on which it was trained, produce a description of the network's hypothesis that is comprehensible yet closely approximates the network's predictive behavior." Such extraction algorithms are useful for experts to verify and cross-check Neural Network systems. Earlier this year, John Launchbury, director of DARPA's Information Innovation Office said, "There's been a lot of hype and bluster about AI." They published their view of AI into "Three Waves," to explain what AI can do, what AI can't do, and where AI is headed. See Fig 3 [18]. We consider the example we have outlined above falls into this "Third Wave of AI."

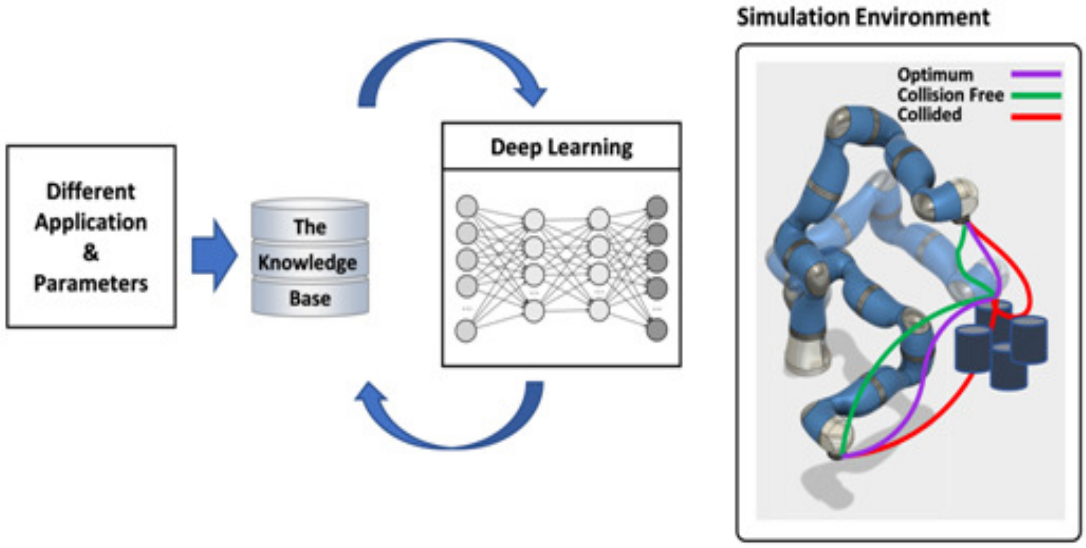


Figure 1: Deep Learning System Applied to Robot Arm Path Planning (Source: CalvIO Inc)

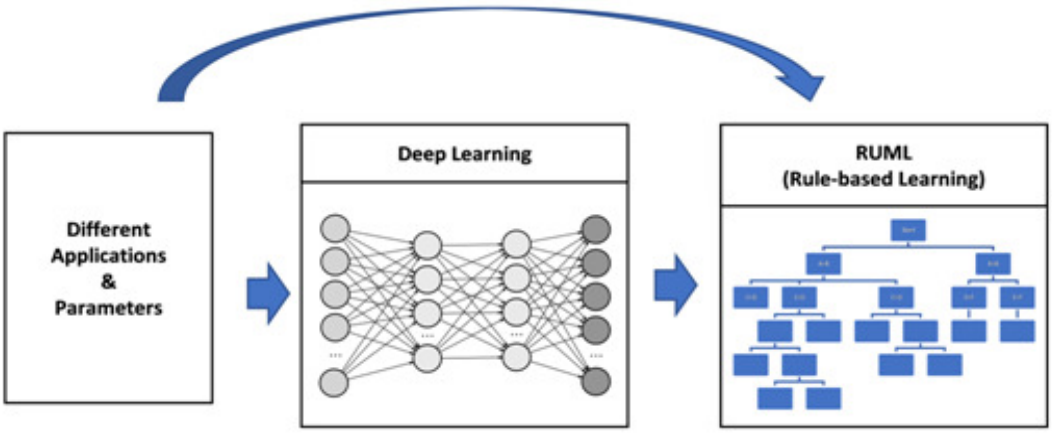


Figure 2: Deep Learning System Rule Extractor (RUML[SM]) (Source: CalvIO Inc)

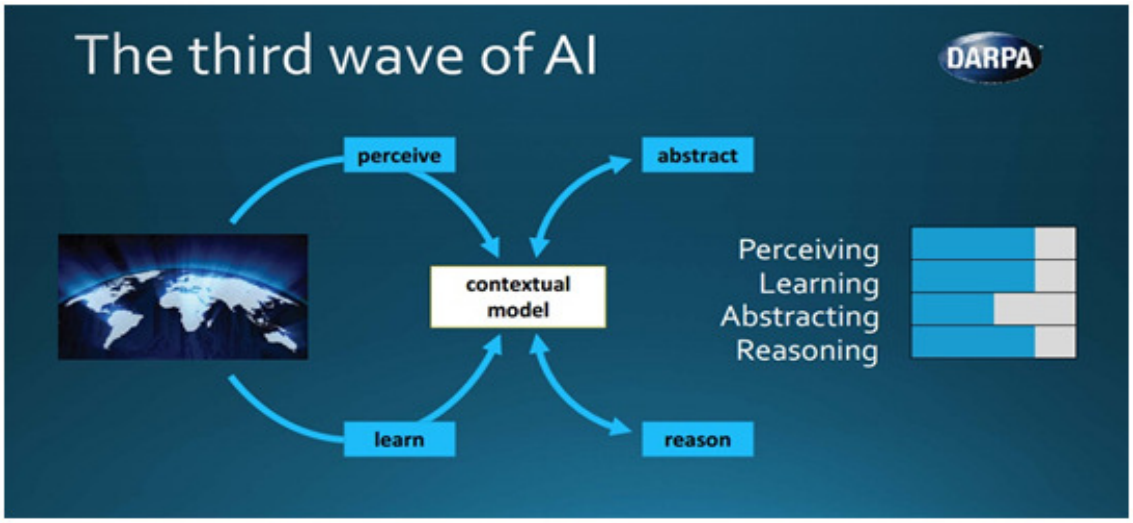


Figure 3: DARPA's Third Wave of AI (Source: DARPA)

3.3 AI Risk Conceptual Analysis applied to the Case Example

1/ Acceptable-Risk Factor: Consent
The use of a robot (in a protective clean room cell) reduces the need for human operator exposure to Biohazards.
2/ Acceptable-Risk Factor: Informed Consent
However, if the AI directing the robot causes breaches of the clean/safe room (e.g., collisions with the cell walls), then what was thought safe might not be. It is necessary to test any black-box defined behavior will not violate these rules.

3/ Acceptable-Risk Factor: The Affected Population
The affected population might not be limited to the factory; conceivably, an extended exposure could cause health and safety threats to those outside, or violations of FDA regulations, etc. Again, the black-box must be validated against the regulations/laws applicable to the domain.
4/ Acceptable-Risk Factor: Step risk versus State risk
Both state and step risks need to be exposed through the black-box rule extraction process.
5/ Acceptable-Risk Factors: Seriousness and Probability
The seriousness of a biohazard breach can be evaluated in principle, but the probability may need testing. This should be done in simulation (to avoid physical exposure) as well as an assessment of the black-box extracted rule-set.
6/ Acceptable-Risk Factors: Who Determines Acceptable Risk?
There are industry bodies that set standards (e.g., GAMP5) as well as government entities that set regulations in this case example (e.g., US FDA).
7/ Acceptable-Risk Factors: The Wild Card: Existential Risk?
If the biohazard agent was severe enough, as might be possible with nuclear materials and/or live chemical/biological agents, then the impacts could be existential, if the AI goes "rogue." The severity relates properly to steps 3 and 5 above. The risk of 'going rogue' is conceivable, but presumably with complete AI automation of the industrial facility, and absent proper safeguards against hacking or rampancy. A solution may involve a software system over-riding ethical kernel that ensures "no harm."

4. CONCLUSIONS

We reviewed the concept of AI Risk and picked a well-known one, i.e., the AI Neural Network Black-box problem. We proceeded to outline a means of structuring a dialog of these "Conceptual AI Risks."

We provided a near-term AI/Robotics case example (Smart Robotics Path Planning for a Pick and Place application for Biohazardous material) and applied the dialog to it; we think the result is an actual beneficial one for highlighting the AI risk concerns and start the process of handling them objectively.

The case example method applies a meta-level/hybrid AI system to "extract" heuristics from the neural network black-boxes (applying a top-down AI system on a bottom-up AI system). The system is based on technology called Really Useful Machine Learning (RUML^{SM}). The approach taken is an example of DARPA's Third Wave of AI. In a sense this process can be considered as a Cybernetic Self-Regulation (extraction of a rationalized model for control).

As such, we believe the resulting techno-philosophy methodology to be a potentially useful early step in the building of tools for conceptualizing and assessing acceptable AI Risk. Further work is needed to develop these concepts, and trial them in real-world applications.

REFERENCES

[1] UNIDIR: The weaponization of increasingly autonomous technologies: considering how Meaningful Human Control might move the discussion forward. UNIDIR Resources, no. 2, 2014. http://unidir.org/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf. Last Referenced 22nd October 2017.

[2] Roff, H.: Monstermind or the doomsday machine? Autonomous cyberwarfare. Duck of Minerva, 13 August 2014. http://duckofminerva.com/2014/08/monstermind-or-the-doomsday-machine-autonomous-cyberwarfare.html. Last Referenced 22nd October 2017.

[3] Clevenger, A.: 'The Terminator conundrum': Pentagon weighs ethics of pairing deadly force, AI. Army Times, 23 January 2016. http://www.armytimes.com/story/defense/policy-budget/budget/2016/01/23/terminator-conundrum-pentagon-weighs-ethics-pairing-deadly-force-ai/79205722/. Last Referenced 22nd October 2017.

[4] Andrew Maynard, "Thinking innovatively about the risks of tech innovation". The Conversation, January 12, 2016. https://theconversation.com/thinking-innovatively-about-the-risks-of-tech-innovation-52934. Last Referenced 22nd October 2017.

[5] Lin, P., Mehlman, M., and Abney, K.: Enhanced warfighters: risk, ethics, and policy. Report funded by the Greenwall Foundation. California Polytechnic State University, San Luis Obispo. 1 January 2013. http://ethics.calpoly.edu/Greenwall_report.pdf. Last Referenced 22nd October 2017.

[6] Abney, K., Lin, P., and Mehlman, M. "Military Neuroenhancement and Risk Assessment" in James Giordano (ed.), Neuroscience and Neurotechnology in National Security and Defense: Practical Considerations, Ethical Concerns (Taylor & Francis Group, 2014)

[7] Lin, P., Mehlman, M., and Abney, K.: Enhanced warfighters: risk, ethics, and policy. Report funded by the Greenwall Foundation. California Polytechnic State University, San Luis Obispo. 1 January 2013. http://ethics.calpoly.edu/Greenwall_report.pdf . Last Referenced 22nd October 2017.

[8] Nick Bostrom, Superintelligence. (Oxford University Press, 2014)

[9] Abney, K., Lin, P., and Mehlman, M. "Military Neuroenhancement and Risk Assessment" in James Giordano (ed.), Neuroscience and Neurotechnology in National Security and Defense: Practical Considerations, Ethical Concerns (Taylor & Francis Group, 2014)

[10] Keith Abney, "Robots and Space Ethics," ch 23 in Robot Ethics 2.0, eds. Lin, P., Jenkins, R., and Abney, K. (Oxford University Press, 2017)

[11] Will Knight, (MIT Press), The Dark Secret at the Heart of AI - No one really knows how the most advanced algorithms do what they do. That could be a problem. April 11, 2017, https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ Last Referenced 22nd October 2017.

[12] Tameru Hailesilassie, (IJCSIS) International Journal of Computer Science and Information Security,Vol. 14, No. 7, July 2016 "Rule Extraction Algorithm for Deep Neural Networks: A Review" https://arxiv.org/ftp/arxiv/papers/1610/1610.05267.pdf Last Referenced 22nd October 2017.

[13] Jan Ruben Zilke, Master Thesis, TUD, "Extracting Rules from Deep Neural Networks" http://www.ke.tu-darmstadt.de/lehre/arbeiten/master/2015/Zilke_Jan.pdf Last Referenced 22nd October 2017.

[14] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, Eric P. Xing, School of Computer Science Carnegie Mellon University, 2016 "Harnessing Deep Neural Networks with Logic Rules" http://www.cs.cmu.edu/~epxing/papers/2016/Hu_etal_ACL16.pdf Last Referenced 22nd October 2017.

[15] M Ciupa “Hybrid Machine Learning Design Based on a Bottom-up/Top-Down Methodology,” US 62/476,068 United States Patent and Trademark Office, 2017.

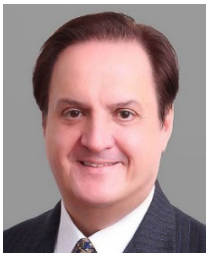
[16] M. Ciupa, "Is AI in Jeopardy? The Need to Under Promise and Over Deliver - The Case for Really Useful Machine Learning," in Computer Science {\&} Information Technology (CS {\&} IT), 2017.

[17] Martin Ciupa, Nicole Tedesco, and Mostafa Ghobadi, “Automating Automation: Master Mentoring Process” 5th International Conference on Artificial Intelligence and Applications (AIAP-2018), Jan 2018, Zurich, Switzerland

[18] Steve Crowe, 21 February 21, 2017, Robotics Trends, “What AI Can and Can’t Do: DARPA’s Realistic View, http://www.roboticstrends.com/article/what_ai_can_and_cant_do_darpas_realistic_view/Artificial_Intelligence

AUTHORS

Martin Ciupa



Martin Ciupa is the CTO of calvIO Inc., a company (associated with the Calvary Robotics group of companies) focused on simplifying the cybernetic interaction between man and machine in the industrial setting. Martin has had a career in both technology, general management and commercial roles at senior levels in North America, Europe and Asia. He has an academic background in Physics and Cybernetics. He has applied AI and Machine learning systems to applications in decision support for Telco, Manufacturing and Financial services sectors and published technical articles in Software, Robotics, AI and related disciplines.

Keith Abney

Keith Abney is senior fellow in the Ethics + Emerging Sciences Group and senior lecturer in the Philosophy Department of California Polytechnic State University - San Luis Obispo. He is co-editor of Robot Ethics (MIT Press) and the newly published Robot Ethics 2.0 (Oxford UP).