ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.015

AI CONCEPTUAL RISK ANALYSIS MATRIX (CRAM^{SM})

MARTIN CIUPA[1] AND KEITH ABNEY[2]
[1]*CTO calvIO Inc., Webster, New York USA*
mciupa@calvIOinc.com
[2]*Senior Lecturer, Cal Poly-SLO, California USA*
kabney@calpoly.edu

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 for designers/implementors of AI-Based systems. We structure this dialog as Conceptual Risk Analysis Matrix (CRAM^{SM}). We look at a topical case example in an actual industrial setting and apply CRAM^{SM}. Conclusions to its efficacy are drawn.

Key Words: AI Risk, Risk Analysis, Dialog Systems.

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, in military robotics serious concern exists over a possible lack of "meaningful human control" [1]. Missing is a similar concern for autonomous AI-controlled cyberattacks 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]. Such issues apply well beyond the military, and demand an analysis of AI risk that also applies to civilian contexts and to risks that do not rise to the level of human extinction.

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. The possible loss of well-being, environmental sustainability, deeply held beliefs, or even a sense of cultural or personal identity should also count. Risk's opposite, safety, should be seen as relative, not absolute: safety in all respects is never 100% guaranteed, so as safety is best understood as relative freedom from a threat of harm, so risk is a relative exposure to such a threat.

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. For example, the development of AGI poses an existential step risk; but, if there is a 'fast takeoff,' any additional state risk of developing AGI may be negligible.

Step risk versus state risk: How shall we determine when state risks are more important than step risks, or vice-versa? If a potential diminishment in a step risk depends on increasing a separate state risk (e.g. slowing down or stopping AGI research that, if successful, would decrease other risks to humanity), how do we decide what to do?

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 or otherwise contraindicated. What magnitude of AI risk is 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? How do we decide on the probability of serious harm that is acceptable, versus moderate harm or mild harm? If a function, is it linear, asymptotic, or other? Is it continuous or not?

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.

2.8 Conceptual Risk Analysis Matrix (CRAM^{SM})

Taking note of items 2.1-2.8 we have formed the following dialog matrix for focusing the dialog of AI Risks in a given use-case.

3. Specific Case Study, Possible Solution, and Risk Analysis

3.1 Specific Case Study

Our case study is applying CRAM^{SM} to the Path Planning of a Robot Arm, in which vials of severe 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 albeit with techno-ethical concerns. The engineering case study has been outlined in Refs [11], [12] and [13].

This path planning problem use case example envisaged here can be described in the following stages of "teaching" the system with a "show & tell" paradigm and AI-based optimization algorithm. It is a process of three stages teaching the system from a basic to advanced level novice to levels of expert proficiency.

Stage 1 proposes Mentor Training, whereby the motion capture of a human expert is captured into a simulation environment.

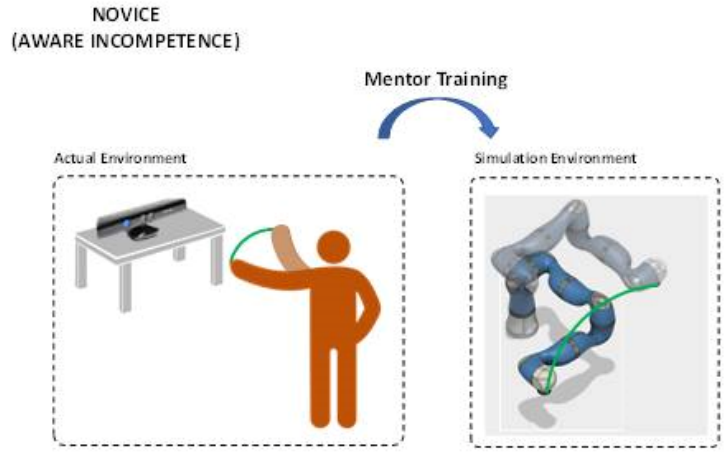
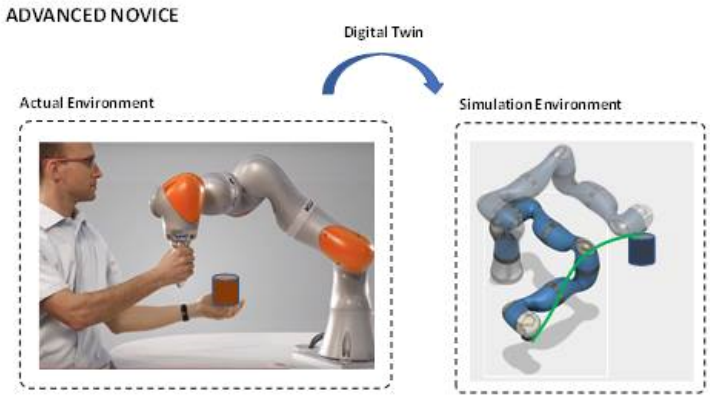


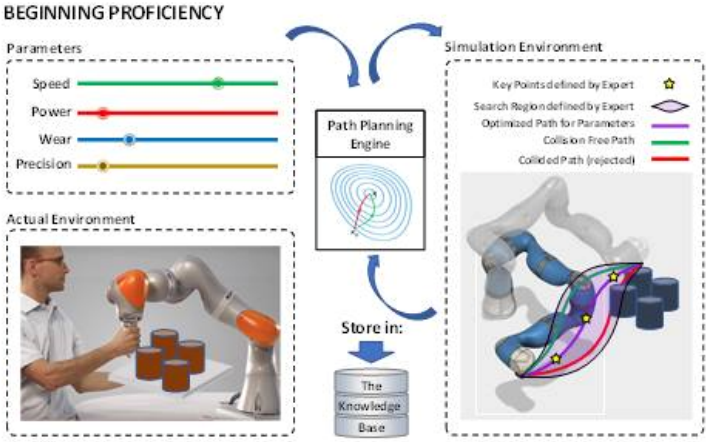
Table 1. Dialog of AI Risks in a given use-case

1/ Acceptable-Risk Factor: Consent
Assess the degree to which consent has been given of the use-case risk
2/ Acceptable-Risk Factor: Informed Consent
Assess the risk of the use-case that parties have potentially not been informed about (or do not understand) the risk potential
3/ Acceptable-Risk Factor: The Affected Population
Assess the risk to the potential Affected Population
4/ Acceptable-Risk Factor: Step risk versus State risk
Assess use case risk in terms of 1. *State Risk (time likely spent in a state that is a cause of risk)* 2. *Step Risk (chance of entering into a new risk, as a consequence of step transitions)*
5/ Acceptable-Risk Factors: Seriousness and Probability
Assess use case's risk in term of an analysis of potential seriousness and probability of occurrence 1. *Seriousness: What (if any) serious risks from AIs are acceptable—and to whom: users, nonusers, the environment, or the AI itself?* 2. *Probability: 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?*
6/ Acceptable-Risk Factors: Who Determines Acceptable Risk?
Assess use case's risk against standards applicable to it. 1. *Good faith subjective standard* 2. *The reasonable-person standard* 3. *Objective standard*
7/ Acceptable-Risk Factors: The Wild Card: Existential Risk?
Assess use case's existential risk that, should it come to pass, might 1. *annihilate Earth-originating intelligent life,* 2. *Or permanently or drastically curtail its potential.*

Stage 2 proposes Digital Twin capture of fine tuning of the Robot arm (a Co-roobotic arm is envisaged in this example that allows for the physical manipulation of the arm).



Stage 3 proposes an AI-Based fine tuning of the path planning by an optimization process based on value parameters (e.g., Speed of path cycle, Power utilization, Wear and Tear, and Precision of movement) an AI search algorithm might seek an optimum minimization of the path plan against a utility metric of these values.



3.2 CRAM[SM] 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. Any personnel entering the clean room cell should have safety training and contracted consent.
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. In this respect personnel in the potential effective area may not be fully informed of the reliability/trust in the system. It is necessary to test any robot behavior in detailed simulation to ensure the path planning algorithms will not likely violate these rules. And personnel potential affected by failures be informed of the extent of the safety testing.
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 system must be validated against the regulations/laws applicable to the domain. The potential affected population should be briefed of the risks.
4/ Acceptable-Risk Factor: Step risk versus State risk
Both state and step risks need to be exposed through the AI testing process and the results passed to stage 5 below.
5/ Acceptable-Risk Factors: Seriousness and Probability
The seriousness of a biohazard breach can be evaluated in principle, but the probability may needs validating in test simulations. Thorough simulation is advocated (to avoid physical exposure) as well as an assessment of the system.
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, in a case of complete AI automation of the industrial facility, and absent proper safeguards against hacking. A solution may involve a software system over-riding ethical kernel that ensures "no harm" and sufficient cybersecurity measures. Ultimately a degree of human oversight may be warranted with the system requiring human authorization for critical procedures. Including the ability to switch off the system.

4. Conclusions

We reviewed the concept of AI Risk and picked a real world industrial problem. We proceeded to outline a means of structuring a dialog we refer to as CRAM[SM]

The AI/Robotics case example (AI-based Path Planning for a Pick and Place application for Biohazardous material) and applied the CRAM[SM] 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.

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] 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.

[12] M Ciupa, N Tedesco, and M Ghobadi, “Automating Automation: Master Mentoring Process” 5th International Conference on Artificial Intelligence and Applications (AIAP-2018), Jan 2018, Zurich, Switzerland
[13] M Ciupa and K Abney, “Conceptualizing AI risk” (AIFU-2018), Melbourne, Australia February, 2018.