Research priorities for robust and beneficial artificial intelligence

January 11, 2015*

Executive Summary: Success in the quest for artificial intelligence has the potential to bring unprecedented benefits to humanity, and it is therefore worthwhile to research how to maximize these benefits while avoiding potential pitfalls. This document gives numerous examples (which should by no means be construed as an exhaustive list) of such worthwhile research aimed at ensuring that AI remains robust and beneficial.

1 Artificial Intelligence Today

Artificial intelligence (AI) research has explored a variety of problems and approaches since its inception, but for the last 20 years or so has been focused on the problems surrounding the construction of *intelligent agents*—systems that perceive and act in some environment. In this context, the criterion for intelligence is related to statistical and economic notions of rationality—colloquially, the ability to make good decisions, plans, or inferences. The adoption of probabilistic representations and statistical learning methods has led to a large degree of integration and cross-fertilization between AI, machine learning, statistics, control theory, neuroscience, and other fields. The establishment of shared theoretical frameworks, combined with the availability of data and processing power, has yielded remarkable successes in various component tasks such as speech recognition, image classification, autonomous vehicles, machine translation, legged locomotion, and question-answering systems.

As capabilities in these areas and others cross the threshold from laboratory research to economically valuable technologies, a virtuous cycle takes hold whereby even small improvements in performance are worth large sums of money, prompting greater investments in research. There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable. Because of the great potential of AI, it is valuable to investigate how to reap its benefits while avoiding potential pitfalls.

The progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI. Such considerations motivated the AAAI 2008–09 Presidential Panel on Long-Term AI Futures [42] and other projects and community efforts on AI impacts. These constitute a significant expansion of the field of AI itself, which up to now has focused largely on techniques that are neutral with respect to purpose. The present document can be viewed as a natural continuation of these efforts, focusing on identifying research directions that can help maximize the societal benefit of AI. This research is by necessity interdisciplinary, because it involves both society and AI. It ranges from economics, law, and philosophy to computer security, formal methods and, of course, various branches of AI itself. The focus is on delivering AI that is *beneficial* to society and *robust* in the sense that the benefits are guaranteed: our AI systems must do what we want them to do.

*The initial version of this document was drafted by Stuart Russell, Daniel Dewey & Max Tegmark, with major input from Janos Kramar & Richard Mallah, and reflects valuable feedback from Anthony Aguirre, Erik Brynjolfsson, Ryan Calo, Tom Dietterich, Dileep George, Bill Hibbard, Demis Hassabis, Eric Horvitz, Leslie Pack Kaelbling, James Manyika, Luke Muehlhauser, Michael Osborne, David Parkes, Heather Roff Perkins, Francesca Rossi, Bart Selman, Murray Shanahan, and many others.

2 Short-term Research Priorities

2.1 Optimizing AI's Economic Impact

The successes of industrial applications of AI, from manufacturing to information services, demonstrate a growing impact on the economy, although there is disagreement about the exact nature of this impact and on how to distinguish between the effects of AI and those of other information technologies. Many economists and computer scientists agree that there is valuable research to be done on how to maximize the economic benefits of AI while mitigating adverse effects, which could include increased inequality and unemployment [50, 12, 25, 26, 71, 52, 48]. Such considerations motivate a range of research directions, spanning areas from economics to psychology. Below are a few examples that should by no means be interpreted as an exhaustive list.

1. **Labor market forecasting:** When and in what order should we expect various jobs to become automated [25]? How will this affect the wages of less skilled workers, creatives, and different kinds of information workers? Some have have argued that AI is likely to greatly increase the overall wealth of humanity as a whole [12]. However, increased automation may push income distribution further towards a power law [13].

2. **Other market disruptions:** Significant parts of the economy, including finance, insurance, actuarial, and many consumer markets, could be susceptible to disruption through the use of AI techniques to learn, model, and predict agent actions. These markets might be indentified by a combination of high complexity and high rewards for navigating that complexity [48].

3. **Policy for managing adverse effects:** What policies could help increasingly automated societies flourish? For example, Brynjolfsson and McAfee [12] explore various policies for incentivizing development of labor-intensive sectors and for using AI-generated wealth to support underemployed populations. What are the pros and cons of interventions such as educational reform, apprenticeship programs, labor-demanding infrastructure projects, and changes to minimum wage law, tax structure, and the social safety net [26]? History provides many examples of subpopulations not needing to work for economic security, ranging from aristocrats in antiquity to many present-day citizens of Qatar. What societal structures and other factors determine whether such populations flourish? Unemployment is not the same as leisure, and there are deep links between unemployment and unhappiness, self-doubt, and isolation [34, 19]; understanding what policies and norms can break these links could significantly improve the median quality of life. Empirical and theoretical research on topics such as the basic income proposal could clarify our options [83, 89].

4. **Economic measures:** It is possible that economic measures such as real GDP per capita do not accurately capture the benefits and detriments of heavily AI-and-automation-based economies, making these metrics unsuitable for policy purposes [50]. Research on improved metrics could be useful for decision-making.

2.2 Law and Ethics Research

The development of systems that embody significant amounts of intelligence and autonomy leads to important legal and ethical questions whose answers impact both producers and consumers of AI technology. These questions span law, professional ethics, and philosophical ethics, and will require expertise from computer scientists, legal experts, policy experts, and ethicists. For example:

1. **Liability and law for autonomous vehicles:** If self-driving cars cut the roughly 40,000 annual US traffic fatalities in half, the car makers might get not 20,000 thank-you notes, but 20,000 lawsuits. In what legal framework can the safety benefits of autonomous vehicles such as drone aircraft and self-driving cars best be realized [85]? Should legal questions about AI be handled by existing (software- and internet-focused) "cyberlaw", or should they be treated separately [14]? In both military and commercial applications, governments will need to decide how best to bring the relevant expertise to bear; for example, a panel or committee of professionals and academics could be created, and Calo has proposed the creation of a Federal Robotics Commission [15].

2. **Machine ethics:** How should an autonomous vehicle trade off, say, a small probability of injury to a human against the near-certainty of a large material cost? How should lawyers, ethicists, and policymakers engage the public on these issues? Should such trade-offs be the subject of national standards?

3. **Autonomous weapons:** Can lethal autonomous weapons be made to comply with humanitarian law [18]? If, as some organizations have suggested, autonomous weapons should be banned, is it possible to develop a precise definition of autonomy for this purpose, and can such a ban practically be enforced? If it is permissible or legal to use lethal autonomous weapons, how should these weapons be integrated into the existing command-and-control structure so that responsibility and liability be distributed, what technical realities and forecasts should inform these questions, and how should "meaningful human control" over weapons be defined [65, 64, 3]? Are autonomous weapons likely to reduce political aversion to conflict, or perhaps result in "accidental" battles or wars [6]? Finally, how can transparency and public discourse best be encouraged on these issues?

4. **Privacy:** How should the ability of AI systems to interpret the data obtained from surveillance cameras, phone lines, emails, *etc.*, interact with the right to privacy? How will privacy risks interact with cyberwarfare? Our ability to take full advantage of the synergy between AI and big data will depend in part on our ability to manage and preserve privacy [47, 1].

5. **Professional ethics:** What role should computer scientists play in the law and ethics of AI development and use? Past and current projects to explore these questions include the AAAI 2008–09 Presidential Panel on Long-Term AI Futures [42], the EPSRC Principles of Robotics [8], and recently-announced programs such as Stanford's One-Hundred Year Study of AI and the AAAI committee on AI impact and ethical issues (chaired by Rossi and Chernova).

From a policy perspective, AI (like any powerful new technology) enables both great new benefits and novel pitfalls to be avoided, and appropriate policies can ensure that we can enjoy the benefits while risks are minimized. This raises policy questions such as these:

1. What is the space of policies worth studying?

2. Which criteria should be used to determine the merits of a policy? Candidates include verifiability of compliance, enforceability, ability to reduce risk, ability to avoid stifling desirable technology development, adoptability, and ability to adapt over time to changing circumstances.

2.3 Computer Science Research for Robust AI

As autonomous systems become more prevalent in society, it becomes increasingly important that they robustly behave as intended. The development of autonomous vehicles, autonomous trading systems, autonomous weapons, *etc.* has therefore stoked interest in high-assurance systems where strong robustness guarantees can be made; Weld and Etzioni have argued that "society will reject autonomous agents unless we have some credible means of making them safe" [88]. Different ways in which an AI system may fail to perform as desired correspond to different areas of robustness research:

1. **Verification:** how to prove that a system satisfies certain desired formal properties. (*"Did I build the system right?"*)

2. **Validity:** how to ensure that a system that meets its formal requirements does not have unwanted behaviors and consequences. (*"Did I build the right system?"*)

3. **Security:** how to prevent intentional manipulation by unauthorized parties.

4. **Control:** how to enable meaningful human control over an AI system after it begins to operate. (*"OK, I built the system wrong, can I fix it?"*)

2.3.1 Verification

By verification, we mean methods that yield high confidence that a system will satisfy a set of formal constraints. When possible, it is desirable for systems in safety-critical situations, e.g. self-driving cars, to be verifiable.

Formal verification of software has advanced significantly in recent years: examples include the *seL4* kernel [43], a complete, general-purpose operating-system kernel that has been mathematically checked against a formal specification to give a strong guarantee against crashes and unsafe operations, and HACMS, DARPA's "clean-slate, formal methods-based approach" to a set of high-assurance software tools [24]. Not only should it be possible to build AI systems on top of verified substrates; it should also be possible to verify the designs of the AI systems themselves, particularly if they follow a "componentized architecture", in which guarantees about individual components can be combined according to their connections to yield properties of the overall system. This mirrors the agent architectures used in Russell and Norvig [68], which separate an agent into distinct modules (predictive models, state estimates, utility functions, policies, learning elements, *etc.*), and has analogues in some formal results on control system designs. Research on richer kinds of agents—for example, agents with layered architectures, anytime components, overlapping deliberative and reactive elements, metalevel control, *etc.*—could contribute to the creation of verifiable agents, but we lack the formal "algebra" to properly define, explore, and rank the space of designs.

Perhaps the most salient difference between verification of traditional software and verification of AI systems is that the correctness of traditional software is defined with respect to a fixed and known machine model, whereas AI systems—especially robots and other embodied systems—operate in environments that are at best partially known by the system designer. In these cases, it may be practical to verify that the system acts correctly given the knowledge that it has, avoiding the problem of modelling the real environment [21]. A lack of design-time knowledge also motivates the use of learning algorithms within the agent software, and verification becomes more difficult: statistical learning theory gives so-called ϵ-δ (probably approximately correct) bounds, mostly for the somewhat unrealistic settings of supervised learning from i.i.d. data and single-agent reinforcement learning with simple architectures and full observability, but even then requiring prohibitively large sample sizes to obtain meaningful guarantees.

Research into methods for making strong statements about the performance of machine learning algorithms and managing computational budget over many different constituent numerical tasks could be improve our abilities in this area, possible extending work on Bayesian quadrature [33, 28]. Work in adaptive control theory [7],the theory of so-called *cyberphysical systems* [57], and verification of hybrid or robotic systems [2, 90] is highly relevant but also faces the same difficulties. And of course all these issues are laid on top of the standard problem of proving that a given software artifact does in fact correctly implement, say, a reinforcement learning algorithm of the intended type. Some work has been done for verifying neural network applications [61, 80, 70] and the notion of *partial programs* [4, 78] allows the designer to impose arbitrary "structural" constraints on behavior, but much remains to be done before it will be possible to have high confidence that a learning agent will learn to satisfy its design criteria in realistic contexts.

2.3.2 Validity

A verification theorem for an agent design has the form, "If environment satisfies assumptions ϕ then behavior satisfies requirements ψ." There are two ways in which a verified agent can, nonetheless, fail to be a beneficial agent in actuality: first, the environmental assumption ϕ is false in the real world, leading to behavior that violates the requirements ψ; second, the system may satisfy the formal requirement ψ but still behave in ways that we find highly undesirable in practice. It may be the case that this undesirability is a consequence of satisfying ψ when ϕ is violated; i.e., had ϕ held the undesirability would not have been manifested; or it may be the case that the requirement ψ is erroneous in itself. Russell and Norvig [68] provide a simple example: if a robot vacuum cleaner is asked to clean up as much dirt as possible, and has an action to dump the contents of its dirt container, it will repeatedly dump and clean up the same dirt. The requirement should focus not on dirt cleaned up but on cleanliness of the floor. Such specification errors are ubiquitous in software verification, where it is commonly observed that writing correct specifications can be harder than writing correct code. Unfortunately, it is not possible to verify the specification: the notions of "beneficial" and "desirable" are not separately made formal, so one cannot straightforwardly prove that satisfying ψ necessarily leads to desirable behavior and a beneficial agent.

In order to build systems that robustly behave well, we of course need to decide what "good behavior" means in each application domain. This ethical question is tied intimately to questions of what engineering techniques are available, how reliable these techniques are, and what trade-offs can be made — all areas where computer science, machine learning, and broader AI expertise is valuable. For example, Wallach and Allen [86] argue that a significant consideration is the computational expense of different behavioral standards (or ethical theories): if a standard cannot be applied efficiently enough to guide behavior in safety-critical situations, then cheaper approximations may be needed. Designing simplified rules – for example, to govern a self-driving car's decisions in critical situations – will likely require expertise from both ethicists and computer scientists. Computational models of ethical reasoning may shed light on questions of computational expense and the viability of reliable ethical reasoning methods [5, 79]; for example, work could further explore the applications of semantic networks for case-based reasoning [49], hierarchical constraint satisfaction [46], or weighted prospective abduction [56] to machine ethics.

2.3.3 Security

Security research can help make AI more robust. As AI systems are used in an increasing number of critical roles, they will take up an increasing proportion of cyber-attack surface area. It is also probable that AI and machine learning techniques will themselves be used in cyber-attacks.

Robustness against exploitation at the low level is closely tied to verifiability and freedom from bugs. For example, the DARPA SAFE program aims to build an integrated hardware-software system with a flexible metadata rule engine, on which can be built memory safety, fault isolation, and other protocols that could improve security by preventing exploitable flaws [20]. Such programs cannot eliminate all security flaws (since verification is only as strong as the assumptions that underly the specification), but could significantly reduce vulnerabilities of the type exploited by the recent "Heartbleed bug" and "Bash Bug". Such systems could be preferentially deployed in safety-critical applications, where the cost of improved security is justified.

At a higher level, research into specific AI and machine learning techniques may become increasingly useful in security. These techniques could be applied to the detection of intrusions [45], analyzing malware [63], or detecting potential exploits in other programs through code analysis [11]. It is not implausible that cyberattack between states and private actors will be a risk factor for harm from near-future AI systems, motivating research on preventing harmful events. As AI systems grow more complex and are networked together, they will have to intelligently manage their trust, motivating research on statistical-behavioral trust establishment [60] and computational reputation models [69].

2.3.4 Control

For certain types of safety-critical AI systems – especially vehicles and weapons platforms – it may be desirable to retain some form of meaningful human control, whether this means a human in the loop, on the loop [35, 55], or some other protocol. In any of these cases, there will be technical work needed in order to ensure that meaningful human control is maintained [22].

Automated vehicles are a test-bed for effective control-granting techniques. The design of systems and protocols for transition between automated navigation and human control is a promising area for further research. Such issues also motivate broader research on how to optimally allocate tasks within human-computer teams, both for identifying situations where control should be transferred, and for applying human judgment efficiently to the highest-value decisions.

3 Long-term research priorities

A frequently discussed long-term goal of some AI researchers is to develop systems that can learn from experience with human-like breadth and surpass human performance in most cognitive tasks, thereby having a major impact on society. If there is a non-negligible probability that these efforts will succeed in the foreseeable future, then additional current research beyond that mentioned in the previous sections will be motivated as exemplified below, to help ensure that the resulting AI will be robust and beneficial.

Assessments of this success probability vary widely between researchers, but few would argue with great confidence that the probability is negligible, given the track record of such predictions. For example, Ernest Rutherford, arguably the greatest nuclear physicist of his time, said in 1933 that nuclear energy was

"moonshine"[1], and Astronomer Royal Richard Woolley called interplanetary travel "utter bilge" in 1956 [62]. Moreover, to justify a modest investment in this AI robustness research, this probability need not be high, merely non-negligible, just as a modest investment in home insurance is justified by a non-negligible probability of the home burning down.

3.1 Verification

Reprising the themes of short-term research, research enabling verifiable low-level software and hardware can eliminate large classes of bugs and problems in general AI systems; if the systems become increasingly powerful and safety-critical, verifiable safety properties will become increasingly valuable. If the theory of extending verifiable properties from components to entire systems is well understood, then even very large systems can enjoy certain kinds of safety guarantees, potentially aided by techniques designed explicitly to handle learning agents and high-level properties. Theoretical research, especially if it is done explicitly with very general and capable AI systems in mind, could be particularly useful.

A related verification research topic that is distinctive to long-term concerns is the verifiability of systems that modify, extend, or improve themselves, possibly many times in succession [27, 84]. Attempting to straightforwardly apply formal verification tools to this more general setting presents new difficulties, including the challenge that a formal system that is sufficiently powerful cannot use formal methods in the obvious way to gain assurance about the accuracy of functionally similar formal systems, on pain of inconsistency via Gödel's incompleteness [23, 87]. It is not yet clear whether or how this problem can be overcome, or whether similar problems will arise with other verification methods of similar strength.

Finally, it is often difficult to actually apply formal verification techniques to physical systems, especially systems that have not been designed with verification in mind. This motivates research pursuing a general theory that links functional specification to physical states of affairs. This type of theory would allow use of formal tools to anticipate and control behaviors of systems that approximate rational agents, alternate designs such as satisficing agents, and systems that cannot be easily described in the standard agent formalism (powerful prediction systems, theorem-provers, limited-purpose science or engineering systems, *etc.*). It may also be that such a theory could allow rigorously demonstrating that systems are constrained from taking certain kinds of actions or performing certain kinds of reasoning.

3.2 Validity

As in the short-term research priorities, validity is concerned with undesirable behaviors that can arise despite a system's formal correctness. In the long term, AI systems might become more powerful and autonomous, in which case failures of validity could carry correspondingly higher costs.

Strong guarantees for machine learning methods, an area we highlighted for short-term validity research, will also be important for long-term safety. To maximize the long-term value of this work, machine learning research might focus on the types of unexpected generalization that would be most problematic for very general and capable AI systems. In particular, it might aim to understand theoretically and practically how learned representations of high-level human concepts could be expected to generalize (or fail to) in radically new contexts [81]. Additionally, if some concepts could be learned reliably, it might be possible to use them to define tasks and constraints that minimize the chances of unintended consequences even when autonomous AI systems become very general and capable. Little work has been done on this topic, which suggests that both theoretical and experimental research may be useful.

Mathematical tools such as formal logic, probability, and decision theory have yielded significant insight into the foundations of reasoning and decision-making. However, there are still many open problems in the foundations of reasoning and decision. Solutions to these problems may make the behavior of very capable systems much more reliable and predictable. Example research topics in this area include reasoning and decision under bounded computational resources à la Horvitz and Russell [40, 66], how to take into account correlations between AI systems' behaviors and those of their environments or of other agents [82, 44, 39, 29, 76], how agents that are embedded in their environments should reason [72, 54], and how to reason about uncertainty over logical consequences of beliefs or other deterministic computations [75, 59]. These topics may benefit from being considered together, since they appear deeply linked [30, 31].

In the long term, it is plausible that we will want to make agents that act autonomously and powerfully across many domains. Explicitly specifying our preferences in broad domains in the style of near-future

[1] "The energy produced by the breaking down of the atom is a very poor kind of thing. Any one who expects a source of power from the transformation of these atoms is talking moonshine" [58].

machine ethics may not be practical, making "aligning" the values of powerful AI systems with our own values and preferences difficult [73, 74]. Consider, for instance, the difficulty of creating a utility function that encompasses an entire body of law; even a literal rendition of the law is far beyond our current capabilities, and would be highly unsatisfactory in practice (since law is written assuming that it will be interpreted and applied in a flexible, case-by-case way). Reinforcement learning raises its own problems: when systems become very capable and general, then an effect similar to Goodhart's Law is likely to occur, in which sophisticated agents attempt to manipulate or directly control their reward signals [9]. This motivates research areas that could improve our ability to engineer systems that can learn or acquire values at run-time. For example, inverse reinforcement learning may offer a viable approach, in which a system infers the preferences of another actor, assumed to be a reinforcement learner itself [67, 51]. Other approaches could use different assumptions about underlying cognitive models of the actor whose preferences are being learned (preference learning, [17]), or could be explicitly inspired by the way humans acquire ethical values. As systems become more capable, more epistemically difficult methods could become viable, suggesting that research on such methods could be useful; for example, Bostrom [9] reviews preliminary work on a variety of methods for specifying goals indirectly.

3.3 Security

It is unclear whether long-term progress in AI will make the overall problem of security easier or harder; on one hand, systems will become increasingly complex in construction and behavior and AI-based cyberattacks may be extremely effective, while on the other hand, the use of AI and machine learning techniques along with significant progress in low-level system reliability may render hardened systems much less vulnerable than today's. From a cryptographic perspective, it appears that this conflict favors defenders over attackers; this may be a reason to pursue effective defense research wholeheartedly.

Although the research topics described in 2.3.3 may become increasingly important in the long term, very general and capable systems will pose distinctive security problems. In particular, if the problems of validity and control are not solved, it may be useful to create "containers" for AI systems that could have undesirable behaviors and consequences in less controlled environments [92]. Both theoretical and practical sides of this question warrant investigation. If the general case of AI containment turns out to be prohibitively difficult, then it may be that designing an AI system and a container in parallel is more successful, allowing the weaknesses and strengths of the design to inform the containment strategy [9]. The design of anomaly detection systems and automated exploit-checkers could be of significant help. Overall, it seems reasonable to expect this additional perspective – defending against attacks from "within" a system as well as from external actors – will raise interesting and profitable questions in the field of computer security.

3.4 Control

It has been argued that very general and capable AI systems operating autonomously to accomplish some task will often be subject to effects that increase the difficulty of maintaining meaningful human control [53, 10, 9, 71]. Research on systems that are not subject to these effects, minimize their impact, or allow for reliable human control could be valuable in preventing undesired consequences, as could work on reliable and secure test-beds for AI systems at a variety of capability levels.

If an AI system is selecting the actions that best allow it to complete a given task, then avoiding conditions that prevent the system from continuing to pursue the task is a natural subgoal [53, 10] (and conversely, seeking unconstrained situations is sometimes a useful heuristic [91]). This could become problematic, however, if we wish to repurpose the system, to deactivate it, or to significantly alter its decision-making process; such a system would rationally avoid these changes. Systems that do not exhibit these behaviors have been termed *corrigible* systems [77], and both theoretical and practical work in this area appears tractable and useful. For example, it may be possible to design utility functions or decision processes so that a system will not try to avoid being shut down or repurposed [77], and theoretical frameworks could be developed to better understand the space of potential systems that avoid undesirable behaviors [36, 38, 37].

It has been argued that another natural subgoal is the acquisition of fungible resources of a variety of kinds: for example, information about the environment, safety from disruption, and improved freedom of action are all instrumentally useful for many tasks [53, 10]. Hammond [32] gives the label *stabilization* to the more general set of cases where "due to the action of the agent, the environment comes to be

better fitted to the agent as time goes on". This type of subgoal could lead to undesired consequences, and a better understanding of the conditions under which resource acquisition or radical stabilization is an optimal strategy (or likely to be selected by a given system) would be useful in mitigating its effects. Potential research topics in this area include "domestic" goals that are limited in scope in some way [9], the effects of large temporal discount rates on resource acquisition strategies, and experimental investigation of simple systems that display these subgoals.

Finally, research on the possibility of superintelligent machines or rapid, sustained self-improvement ("intelligence explosion") has been highlighted by past and current projects on the future of AI as potentially valuable to the project of maintaining reliable control in the long term. The the AAAI 2008–09 Presidential Panel on Long-Term AI Futures' "Subgroup on Pace, Concerns, and Control" stated that

> There was overall skepticism about the prospect of an intelligence explosion... Nevertheless, there was a shared sense that additional research would be valuable on methods for understanding and verifying the range of behaviors of complex computational systems to minimize unexpected outcomes. Some panelists recommended that more research needs to be done to better define "intelligence explosion," and also to better formulate different classes of such accelerating intelligences. Technical work would likely lead to enhanced understanding of the likelihood of such phenomena, and the nature, risks, and overall outcomes associated with different conceived variants [42].

Stanford's One-Hundred Year Study of Artificial Intelligence includes "Loss of Control of AI systems" as an area of study, specifically highlighting concerns over the possibility that

> we could one day lose control of AI systems via the rise of superintelligences that do not act in accordance with human wishes – and that such powerful systems would threaten humanity. Are such dystopic outcomes possible? If so, how might these situations arise? ...What kind of investments in research should be made to better understand and to address the possibility of the rise of a dangerous superintelligence or the occurrence of an "intelligence explosion"? [41]

Research in this area could include any of the long-term research priorities listed above, as well as theoretical and forecasting work on intelligence explosion and superintelligence [16, 9], and could extend or critique existing approaches begun by groups such as the Machine Intelligence Research Institute [74].

4 Conclusion

In summary, success in the quest for artificial intelligence has the potential to bring unprecedented benefits to humanity, and it is therefore worthwhile to research how to maximize these benefits while avoiding potential pitfalls. This document has given numerous examples (which should by no means be construed as an exhaustive list) of such worthwhile research aimed at ensuring that AI remains robust and beneficial, and aligned with human interests.

References

[1] Rakesh Agrawal and Ramakrishnan Srikant. "Privacy-preserving data mining". In: *ACM Sigmod Record* 29.2 (2000), pp. 439–450.

[2] Rajeev Alur. "Formal verification of hybrid systems". In: *Embedded Software (EMSOFT), 2011 Proceedings of the International Conference on*. IEEE. 2011, pp. 273–278.

[3] Kenneth Anderson, Daniel Reisner, and Matthew C Waxman. "Adapting the Law of Armed Conflict to Autonomous Weapon Systems". In: (2014).

[4] David Andre and Stuart J Russell. "State abstraction for programmable reinforcement learning agents". In: *Eighteenth national conference on Artificial intelligence*. American Association for Artificial Intelligence. 2002, pp. 119–125.

[5] Peter M Asaro. "What should we want from a robot ethic?" In: *International Review of Information Ethics* 6.12 (2006), pp. 9–16.

[6] Peter Asaro. "How just could a robot war be". In: *Current issues in computing and philosophy* (2008), pp. 50–64.

[7] Karl J Åström and Björn Wittenmark. *Adaptive control*. Courier Dover Publications, 2013.

[8] M Boden et al. "Principles of robotics". In: *The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC). web publication* (2011).

[9] Nick Bostrom. *Superintelligence: Paths, dangers, strategies*. Oxford University Press, 2014.

[10] Nick Bostrom. "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents". In: *Minds and Machines* 22.2 (2012), pp. 71–85.

[11] Yuriy Brun and Michael D Ernst. "Finding latent code errors via machine learning over program executions". In: *Proceedings of the 26th International Conference on Software Engineering*. IEEE Computer Society. 2004, pp. 480–490.

[12] Erik Brynjolfsson and Andrew McAfee. *The second machine age: work, progress, and prosperity in a time of brilliant technologies*. W.W. Norton & Company, 2014.

[13] Erik Brynjolfsson, Andrew McAfee, and Michael Spence. "Labor, Capital, and Ideas in the Power Law Economy". In: *Foreign Aff.* 93 (2014), p. 44.

[14] Ryan Calo. "Robotics and the New Cyberlaw". In: *Available at SSRN 2402972* (2014).

[15] Ryan Calo. "The Case for a Federal Robotics Commission". In: *Available at SSRN 2529151* (2014).

[16] David Chalmers. "The singularity: A philosophical analysis". In: *Journal of Consciousness Studies* 17.9-10 (2010), pp. 7–65.

[17] Wei Chu and Zoubin Ghahramani. "Preference learning with Gaussian processes". In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 2005, pp. 137–144.

[18] Robin R Churchill and Geir Ulfstein. "Autonomous institutional arrangements in multilateral environmental agreements: a little-noticed phenomenon in international law". In: *American Journal of International Law* (2000), pp. 623–659.

[19] Andrew E Clark and Andrew J Oswald. "Unhappiness and unemployment". In: *The Economic Journal* (1994), pp. 648–659.

[20] André DeHon et al. "Preliminary design of the SAFE platform". In: *Proceedings of the 6th Workshop on Programming Languages and Operating Systems*. ACM. 2011, p. 4.

[21] Louise A Dennis et al. "Practical Verification of Decision-Making in Agent-Based Autonomous Systems". In: *arXiv preprint arXiv:1310.2431* (2013).

[22] United Nations Institute for Disarmament Research. *The Weaponization of Increasingly Autonomous Technologies: Implications for Security and Arms Control*. UNIDIR, 2014.

[23] Benja Fallenstein and Nate Soares. *Vingean Reflection: Reliable Reasoning for Self-Modifying Agents*. Tech. rep. Machine Intelligence Research Institute, 2014. URL: `https://intelligence.org/files/VingeanReflection.pdf`.

[24] Kathleen Fisher. "HACMS: high assurance cyber military systems". In: *Proceedings of the 2012 ACM conference on high integrity language technology*. ACM. 2012, pp. 51–52.

[25] Carl Frey and Michael Osborne. *The future of employment: how susceptible are jobs to computerisation?* Working Paper. Oxford Martin School, 2013.

[26] Edward L Glaeser. "Secular joblessness". In: *Secular Stagnation: Facts, Causes and Cures* (2014), p. 69.

[27] Irving John Good. "Speculations concerning the first ultraintelligent machine". In: *Advances in computers* 6.31 (1965), p. 88.

[28] Tom Gunter et al. "Sampling for inference in probabilistic models with fast Bayesian quadrature". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2789–2797.

[29] Joseph Y Halpern and Rafael Pass. "Game theory with translucent players". In: *arXiv preprint arXiv:1308.3778* (2013).

[30] Joseph Y Halpern and Rafael Pass. "I don't want to think about it now: Decision theory with costly computation". In: *arXiv preprint arXiv:1106.2657* (2011).

[31] Joseph Y Halpern, Rafael Pass, and Lior Seeman. "Decision Theory with Resource-Bounded Agents". In: *Topics in cognitive science* 6.2 (2014), pp. 245–257.

[32] Kristian J Hammond, Timothy M Converse, and Joshua W Grass. "The stabilization of environments". In: *Artificial Intelligence* 72.1 (1995), pp. 305–327.

[33] Philipp Hennig and Martin Kiefel. "Quasi-Newton methods: A new direction". In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 843–865.

[34] Clemens Hetschko, Andreas Knabe, and Ronnie Schöb. "Changing identity: Retiring from unemployment". In: *The Economic Journal* 124.575 (2014), pp. 149–166.

[35] Henry Hexmoor, Brian McLaughlan, and Gaurav Tuli. "Natural human role in supervising complex control systems". In: *Journal of Experimental & Theoretical Artificial Intelligence* 21.1 (2009), pp. 59–77.

[36] Bill Hibbard. "Avoiding unintended AI behaviors". In: *Artificial General Intelligence*. Springer, 2012, pp. 107–116.

[37] Bill Hibbard. *Ethical Artificial Intelligence*. 2014. URL: `arxiv.org/abs/1411.1373`.

[38] Bill Hibbard. "Self-Modeling Agents and Reward Generator Corruption". In: *AAAI-15 Workshop on AI and Ethics*. 2015.

[39] Daniel Hintze. "Problem Class Dominance in Predictive Dilemmas". Honors Thesis. Arizona State University, 2014.

[40] Eric J Horvitz. "Reasoning about beliefs and actions under computational resource constraints". In: *Third AAAI Workshop on Uncertainty in Artificial Intelligence*. 1987, pp. 429–444.

[41] Eric Horvitz. *One-Hundred Year Study of Artificial Intelligence: Reflections and Framing*. White paper. Stanford University, 2014. URL: `https://stanford.app.box.com/s/266hrhww2l3gjoy9euar`.

[42] Eric Horvitz and Bart Selman. *Interim Report from the Panel Chairs*. AAAI Presidential Panel on Long Term AI Futures. 2009. URL: `https://www.aaai.org/Organization/Panel/panel-note.pdf`.

[43] Gerwin Klein et al. "seL4: Formal verification of an OS kernel". In: *Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles*. ACM. 2009, pp. 207–220.

[44] Patrick LaVictoire et al. "Program Equilibrium in the Prisoner's Dilemma via Löb's Theorem". In: *AAAI Multiagent Interaction without Prior Coordination workshop*. 2014.

[45] Terran D Lane. "Machine learning techniques for the computer security domain of anomaly detection". PhD thesis. Purdue University, 2000.

[46] Alan K Mackworth. "Agents, bodies, constraints, dynamics, and evolution". In: *AI Magazine* 30.1 (2009), p. 7.

[47] James Manyika et al. "Big data: The next frontier for innovation, competition, and productivity". In: (2011).

[48] James Manyika et al. *Disruptive technologies: Advances that will transform life, business, and the global economy*. Vol. 180. McKinsey Global Institute San Francisco, CA, 2013.

[49] Bruce M McLaren. "Computational models of ethical reasoning: Challenges, initial steps, and future directions". In: *Intelligent Systems, IEEE* 21.4 (2006), pp. 29–37.

[50] Joel Mokyr. "Secular stagnation? Not in your life". In: *Secular Stagnation: Facts, Causes and Cures* (2014), p. 83.

[51] Andrew Y Ng and Stuart Russell. "Algorithms for Inverse Reinforcement Learning". In: *in Proc. 17th International Conf. on Machine Learning*. Citeseer. 2000.

[52] Nils J Nilsson. "Artificial intelligence, employment, and income". In: *AI Magazine* 5.2 (1984), p. 5.

[53] Stephen M Omohundro. *The nature of self-improving artificial intelligence*. Presented at Singularity Summit 2007.

[54] Laurent Orseau and Mark Ring. "Space-Time embedded intelligence". In: *Artificial General Intelligence*. Springer, 2012, pp. 209–218.

[55] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. "A model for types and levels of human interaction with automation". In: *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 30.3 (2000), pp. 286–297.

[56] Luís Moniz Pereira and Ari Saptawijaya. "Modelling morality with prospective logic". In: *Progress in Artificial Intelligence*. Springer, 2007, pp. 99–111.

[57] Andr Platzer. *Logical analysis of hybrid systems: proving theorems for complex dynamics*. Springer Publishing Company, Incorporated, 2010.

[58] Associated Press. "Atom-Powered World Absurd, Scientists Told". In: *New York Herald Tribune* (). September 12, 1933, p. 1.

[59] *Probabilistic Numerics*. http://probabilistic-numerics.org. Accessed: 27 November 2014.

[60] Matthew J Probst and Sneha Kumar Kasera. "Statistical trust establishment in wireless sensor networks". In: *Parallel and Distributed Systems, 2007 International Conference on*. Vol. 2. IEEE. 2007, pp. 1–8.

[61] Luca Pulina and Armando Tacchella. "An abstraction-refinement approach to verification of artificial neural networks". In: *Computer Aided Verification*. Springer. 2010, pp. 243–257.

[62] Reuters. "Space Travel 'Utter Bilge'". In: *The Ottawa Citizen* (). January 3, 1956, p. 1. URL: http://news.google.com/newspapers?id=ddgxAAAAIBAJ&sjid=1eMFAAAAIBAJ&pg=3254%2C7126.

[63] Konrad Rieck et al. "Automatic analysis of malware behavior using machine learning". In: *Journal of Computer Security* 19.4 (2011), pp. 639–668.

[64] Heather M Roff. "Responsibility, liability, and lethal autonomous robots". In: *Routledge Handbook of Ethics and War: Just War Theory in the 21st Century* (2013), p. 352.

[65] Heather M Roff. "The Strategic Robot Problem: Lethal Autonomous Weapons in War". In: *Journal of Military Ethics* 13.3 (2014).

[66] Stuart J Russell and Devika Subramanian. "Provably bounded-optimal agents". In: *Journal of Artificial Intelligence Research* (1995), pp. 1–36.

[67] Stuart Russell. "Learning agents for uncertain environments". In: *Proceedings of the eleventh annual conference on Computational learning theory*. ACM. 1998, pp. 101–103.

[68] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd. Pearson, 2010.

[69] Jordi Sabater and Carles Sierra. "Review on computational trust and reputation models". In: *Artificial intelligence review* 24.1 (2005), pp. 33–60.

[70] Johann M Schumann and Yan Liu. *Applications of neural networks in high assurance systems*. Springer, 2010.

[71] Murray Shanahan. *The Technological Singularity*. Forthcoming. MIT Press, 2015.

[72] Nate Soares. *Formalizing Two Problems of Realistic World-Models*. Tech. rep. Machine Intelligence Research Institute, 2014. URL: https://intelligence.org/files/RealisticWorldModels.pdf.

[73] Nate Soares. *The Value Learning Problem*. Tech. rep. Machine Intelligence Research Institute, 2014. URL: https://intelligence.org/files/ValueLearningProblem.pdf.

[74] Nate Soares and Benja Fallenstein. *Aligning Superintelligence with Human Interests: A Technical Research Agenda*. Tech. rep. Machine Intelligence Research Institute, 2014. URL: http://intelligence.org/files/TechnicalAgenda.pdf.

[75] Nate Soares and Benja Fallenstein. *Questions of Reasoning Under Logical Uncertainty*. Tech. rep. URL: http://intelligence.org/files/QuestionsLogicalUncertainty.pdf. Machine Intelligence Research Institute, 2014.

[76] Nate Soares and Benja Fallenstein. *Toward Idealized Decision Theory*. Tech. rep. URL: https://intelligence.org/files/TowardIdealizedDecisionTheory.pdf. Machine Intelligence Research Institute, 2014.

[77] Nate Soares et al. "Corrigibility". In: *AAAI-15 Workshop on AI and Ethics*. 2015. URL: http://intelligence.org/files/Corrigibility.pdf.

[78] Diana F Spears. "Assuring the behavior of adaptive agents". In: *Agent technology from a formal perspective*. Springer, 2006, pp. 227–257.

[79] John P Sullins. "Introduction: Open questions in roboethics". In: *Philosophy & Technology* 24.3 (2011), pp. 233–238.

[80] Brian J. (Ed.) Taylor. *Methods and Procedures for the Verification and Validation of Artificial Neural Networks*. Springer, 2006.

[81] Max Tegmark. "Friendly Artificial Intelligence: the Physics Challenge". In: *AAAI-15 Workshop on AI and Ethics*. 2015. URL: http://arxiv.org/pdf/1409.0813.pdf.

[82] Moshe Tennenholtz. "Program equilibrium". In: *Games and Economic Behavior* 49.2 (2004), pp. 363–373.

[83] Philippe Van Parijs et al. *Arguing for Basic Income. Ethical foundations for a radical reform*. Verso, 1992.

[84] Vernor Vinge. "The coming technological singularity". In: *VISION-21 Symposium, NASA Lewis Research Center and the Ohio Aerospace Institute*. NASA CP-10129. 1993. URL: http://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html.

[85] David C Vladeck. "Machines without Principals: Liability Rules and Artificial Intelligence". In: *Wash. L. Rev.* 89 (2014), p. 117.

[86] Wendell Wallach and Colin Allen. *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

[87] Nik Weaver. *Paradoxes of rational agency and formal systems that verify their own soundness*. Preprint. URL: http://arxiv.org/pdf/1312.3626.pdf.

[88] Daniel Weld and Oren Etzioni. "The first law of robotics (a call to arms)". In: *AAAI*. Vol. 94. 1994, pp. 1042–1047.

[89] Karl Widerquist et al. "Basic income: an anthology of contemporary research". In: (2013).

[90] Alan FT Winfield, Christian Blum, and Wenguo Liu. "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection". In: *Advances in Autonomous Robotics Systems*. Springer, 2014, pp. 85–96.

[91] AD Wissner-Gross and CE Freer. "Causal entropic forces". In: *Physical review letters* 110.16 (2013), p. 168702.

[92] Roman Yampolskiy. "Leakproofing the Singularity: Artificial Intelligence Confinement Problem". In: *Journal of Consciousness Studies* 19.1-2 (2012), pp. 1–2.