Research Priorities for Robust and Beneficial Artificial Intelligence

Citation: Russell, Stuart; Dewey, Daniel and Tegmark, Max Erik. "Research Priorities for Robust and Beneficial Artificial Intelligence." AI Magazine 36, no. 4 (December 2015): 105-114. © 2015 Association for the Advancement of Artificial Intelligence.

As Published: https://aaai.org/ojs/index.php/aimagazine/article/view/2577/2521

Publisher: Association for the Advancement of Artificial Intelligence

Persistent URL: http://hdl.handle.net/1721.1/108478

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Research Priorities for Robust and Beneficial Artificial Intelligence

Stuart Russell, Daniel Dewey, Max Tegmark

■ *Success in the quest for artificial intelligence has the potential to bring unprecedented benefits to humanity, and it is therefore worthwhile to investigate how to maximize these benefits while avoiding potential pitfalls. This article 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.*

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 have significant economic value, 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 is not unfathomable. Because of the great potential of AI, it is valuable to investigate how to reap its benefits while avoiding potential pitfalls.

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 (Horvitz and Selman 2009) and other projects and community efforts on AI's future 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.

This article was drafted with input from the attendees of the 2015 conference The Future of AI: Opportunities and Challenges (see Acknowledgements), and was the basis for an open letter that has collected nearly 7000 signatures in support of the research priorities outlined here.

Short-Term Research Priorities

Short-term research priorities including optimizing AI's economic impact, research in law and ethics, and computer science research for robust AI. In this section, each of these priorities will, in turn, be discussed.

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 (Mokyr 2014; Brynjolfsson and McAfee 2014; Frey and Osborne 2013; Glaeser 2014; Shanahan 2015; Nilsson 1984; Manyika et al. 2013). Such considerations motivate a range of research directions, spanning areas from economics to psychology. Examples include the following.

Labor Market Forecasting:

When and in what order should we expect various jobs to become automated (Frey and Osborne 2013)? How will this affect the wages of less skilled workers, the creative professions, and various kinds of information workers? Some have have argued that AI is likely to greatly increase the overall wealth of humanity as a whole (Brynjolfsson and McAfee 2014). However, increased automation may push income distribution further towards a power law (Brynjolfsson, McAfee, and Spence 2014), and the resulting disparity may fall disproportionately along lines of race, class, and gender; research anticipating the economic and societal impact of such disparity could be useful.

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 human and market behaviors. These markets might be identified by a combination of high complexity and high rewards for navigating that complexity (Manyika et al. 2013).

Policy for Managing Adverse Effects

What policies could help increasingly automated societies flourish? For example, Brynjolfsson and McAfee (2014) 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 (Glaeser 2014)? 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 (Hetschko, Knabe, and Schöb 2014; Clark and Oswald 1994); 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 (Van Parijs 1992; Widerquist et al. 2013).

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 (Mokyr 2014). Research on improved metrics could be useful for decision making.

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 affect both producers and consumers of AI technology. These questions span law, public policy, professional ethics, and philosophical ethics, and will require expertise from computer scientists, legal experts, political scientists, and ethicists. For example:

Liability and Law for Autonomous Vehicles

If self-driving cars cut the roughly 40,000 annual U.S. 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 (Vladeck 2014)? Should legal questions about AI be handled by existing (software- and Internet-focused) cyberlaw, or should they be treated separately (Calo 2014b)? 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 (Calo 2014a).

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?

Autonomous Weapons

Can lethal autonomous weapons be made to comply with humanitarian law (Churchill and Ulfstein 2000)? If, as some organizations have suggested, autonomous weapons should be banned (Docherty 2012), 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 remain associated with specific human actors? What technical realities and forecasts should inform these questions, and how should meaningful human control over weapons be defined (Roff 2013, 2014; Anderson, Reisner, and Waxman 2014)? Are autonomous weapons likely to reduce political aversion to conflict, or perhaps result in accidental battles or wars (Asaro 2008)? Would such weapons become the tool of choice for oppressors or terrorists? Finally, how can transparency and public discourse best be encouraged on these issues?

Privacy

How should the ability of AI systems to interpret the data obtained from surveillance cameras, phone lines, emails, and so on, interact with the right to privacy? How will privacy risks interact with cybersecurity and cyberwarfare (Singer and Friedman 2014)? 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 (Manyika et al. 2011; Agrawal and Srikant 2000).

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 (Horvitz and Selman 2009), the EPSRC Principles of Robotics (Boden et al. 2011), and recently announced programs such as Stanford's One-Hundred Year Study of AI and the AAAI Committee on AI Impact and Ethical Issues.

Policy Questions

From a public 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 (1) What is the space of policies worth studying, and how might they be enacted? (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, likelihood of being adoped, and ability to adapt over time to changing circumstances.

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, and so on, has therefore stoked interest in high-assurance systems where strong robustness guarantees can be made; Weld and Etzioni (1994) have argued that "society will reject autonomous agents unless we have some credible means of making them safe." Different ways in which an AI system may fail to perform as desired correspond to different areas of robustness research:

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

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?)

Security: How to prevent intentional manipulation by unauthorized parties.

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?)

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, for example, self-driving cars, to be verifiable.

Formal verification of software has advanced significantly in recent years: examples include the seL4 kernel (Klein et al. 2009), 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 (Fisher 2012). 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 (2010), which separate an agent into distinct modules (predictive models, state estimates, utility functions, policies, learning elements, and others), 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, and so on — 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 (Dennis et al. 2013). 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 $\varepsilon - \delta$ (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.

Work in adaptive control theory (Åström and Wittenmark 2013), the theory of so-called cyberphysical systems (Platzer 2010), and verification of hybrid or robotic systems (Alur 2011; Winfield, Blum, and Liu 2014) 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 on verifying neural network applications (Pulina and Tacchella 2010; Taylor 2006; Schumann and Liu 2010) and the notion of partial programs (Andre and Russell 2002; Spears 2006) 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.

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; that is, 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 (2010) 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 (2008) 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 (Asaro 2006, Sullins 2011).

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 cyberattack surface area. It is also probable that AI and machine-learning techniques will themselves be used in cyberattacks.

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 (DeHon et al. 2011). 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 and Bash bugs. 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 (Lane 2000), analyzing malware (Rieck et al. 2011), or detecting potential exploits in other programs through code analysis (Brun and Ernst 2004). 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 (Probst and Kasera 2007) and computational reputation models (Sabater and Sierra 2005).

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 (Hexmoor, McLaughlan, and Tuli 2009; Parasuraman, Sheridan, and Wickens 2000), 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 (UNIDIR 2014).

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.

Long-Term Research Priorities

A frequently discussed long-term goal of some AI researchers is to develop systems that can learn from experience with humanlike 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 next, 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 — less than 24 hours before Szilard's invention of the nuclear chain reaction — that nuclear energy was "moonshine" (Press 1933), and astronomer Royal Richard Woolley called interplanetary travel "utter bilge" in 1956 (Reuters 1956). Moreover, to justify a modest investment in this AI robustness research, this probability need not be high, merely nonnegligible, just as a modest investment in home insurance is justified by a nonnegligible probability of the home burning down.

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 such 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 (Good 1965, Vinge 1993). 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 through Gödel's incompleteness (Fallenstein and Soares 2014; Weaver 2013). 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, and so on). It may also be that such a theory could allow rigorous demonstrations that systems are constrained from taking certain kinds of actions or performing certain kinds of reasoning.

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 (Tegmark 2015). 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 (Horvitz 1987; Russell and Subramanian 1995), how to take into account correlations between AI systems' behaviors and those of their environments or of other agents (Tennenholtz 2004; LaVictoire et al. 2014; Hintze 2014; Halpern and Pass 2013; Soares and Fallenstein 2014c), how agents that are embedded in their environments should reason (Soares 2014a; Orseau and Ring 2012), and how to reason about uncertainty over logical consequences of beliefs or other deterministic computations (Soares and Fallenstein 2014b). These topics may benefit from being considered together, since they appear deeply linked (Halpern and Pass 2011; Halpern, Pass, and Seeman 2014).

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 machine ethics may not be practical, making aligning the values of powerful AI systems with our own values and preferences difficult (Soares 2014b, Soares and Fallenstein 2014a).

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 by humans who, presumably, already embody the background value systems that artificial agents may lack). 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 (Bostrom 2014). 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 rational or nearly rational actor by observing its behavior (Russell 1998, Ng and Russell 2000). Other approaches could use different assumptions about underlying cognitive models of the actor whose preferences are being learned (Chu and Ghahramani 2005), 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 (2014) reviews preliminary work on a variety of methods for specifying goals indirectly.

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 topics described in the near-term security research section earlier 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 (Yampolskiy 2012). 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 (Bostrom 2014). 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.

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 (Omohundro 2007; Bostrom 2012, 2014; Shanahan 2015). 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 (Omohundro 2007, Bostrom 2012) (and conversely, seeking unconstrained situations is sometimes a useful heuristic [Wissner-Gross and Freer 2013]). 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 (Soares et al. 2015), 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 (Soares et al. 2015), and theoretical frameworks could be developed to better understand the space of potential systems that avoid undesirable behaviors (Hibbard 2012, 2014, 2015).

It has been argued that another natural subgoal for AI systems pursuing a given goal 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 (Omohundro 2007, Bostrom 2012). Hammond et al. (1995) give 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 (Bostrom 2014), 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 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 (Horvitz and Selman 2009).

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"? (Horvitz 2014)

Research in this area could include any of the long-term research priorities listed previously, as well as theoretical and forecasting work on intelligence explosion and superintelligence (Chalmers 2010, Bostrom 2014), and could extend or critique existing approaches begun by groups such as the Machine Intelligence Research Institute (Soares and Fallenstein 2014a).

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. The research agenda outlined in this paper, and the concerns that motivate it, have been called anti-AI, but we vigorously contest this characterization. It seems self-evident that the growing capabilities of AI are leading to an increased potential for impact on human society. It is the duty of AI researchers to ensure that the future impact is beneficial. We believe that this is possible, and hope that this research agenda provides a helpful step in the right direction.

Acknowledgements

The initial version of this document was drafted with major input from Janos Kramar and Richard Mallah, and reflects valuable feedback from Anthony Aguirre, Erik Brynjolfsson, Ryan Calo, Meia Chita-Tegmark, Tom Dietterich, Dileep George, Bill Hibbard, Demis Hassabis, Eric Horvitz, Leslie Pack Kaelbling, James Manyika, Luke Muehlhauser, Michael Osborne, David Parkes, Heather Roff, Francesca Rossi, Bart Selman, Murray Shanahan, and many others. The authors are also grateful to Serkan Cabi and David Stanley for help with manuscript editing and formatting.

References

Agrawal, R., and Srikant, R. 2000. Privacy-Preserving Data Mining. *ACM Sigmod Record* 29(2): 439–450. dx.doi.org/10.1145/335191.335438

Alur, R. 2011. Formal Verification of Hybrid Systems. In *Proceedings of the 2011 IEEE International Conference on Embedded Software (EMSOFT),* 273–278. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1145/2038642.2038685

Anderson, K.; Reisner, D.; and Waxman, M. C. 2014. Adapting the Law of Armed Conflict to Autonomous Weapon Systems. *International Law Studies* 90: 386–411.

Andre, D., and Russell, S. J. 2002. State Abstraction For Programmable Reinforcement Learning Agents. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence,* 119–125. Menlo Park, CA: AAAI Press.

Asaro, P. 2008. How Just Could a Robot War Be? In *Current Issues in Computing and Philosophy,* ed. K. W. Adam Briggle and P. A. E. Brey , 50–64. Amsterdam: IOS Press.

Asaro, P. M. 2006. What Should We Want from a Robot Ethic? I*nternational Review of Information Ethics* 6(12): 9–16.

Åström, K. J., and Wittenmark, B. 2013. *Adaptive Control*. Mineola, NY: Courier Dover Publications.

Boden, M.; Bryson, J.; Caldwell, D.; Dautenhahn, K.; Edwards, L.; Kember, S.; Newman, P.; Parry, V.; Pegman, G.; Rodden, T.; Sorell, T.; Wallis, M.; WHitby, B.; Winfield, A.; and Parry, V. 2011. *Principles of Robotics*. Swindon, UK: Engineering and Physical Sciences Research Council. (www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics)

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press.

Bostrom, N. 2012. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines* 22(2): 71–85. dx.doi.org/10.1007/s11023-012-9281-3

Brun, Y., and Ernst, M. D. 2004. Finding Latent Code Errors Via Machine Learning over Program Executions. In *Proceedings of the 26th International Conference on Software Engineering,* 480–495. Los Alamitos, CA: IEEE Computer Society. dx.doi.org/10.1109/ICSE.2004.1317470

Brynjolfsson, E., and McAfee, A. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W.W. Norton & Company.

Brynjolfsson, E.; McAfee, A.; and Spence, M. 2014. Labor, Capital, and Ideas in the Power Law Economy. *Foreign Affairs* 93(4): 44.

Calo, R. 2014a. The Case for a Federal Robotics Commission. Brookings Institution Report (May 2014). Washington, DC: Brookings Institution.

Calo, R. 2014b. Robotics and the Lessons of Cyberlaw. University of Washington School of Law Legal Studies Research Paper No. 2014.08. Seattle, WA: University of Washington.

Chalmers, D. 2010. The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies* 17(9–10): 7–65.

Chu, W., and Ghahramani, Z. 2005. Preference Learning with Gaussian Processes. In *Proceedings of the 22nd International Conference on Machine Learning,* 137–144. New York: Association for Computing Machinery. dx.doi.org/10.1145/1102351.1102369

Churchill, R. R., and Ulfstein, G. 2000. Autonomous Institutional Arrangements in Multilateral Environmental Agreements: A Little-Noticed Phenomenon in International Law. *American Journal of International Law* 94(4): 623–659. dx.doi.org/10.2307/2589775

Clark, A. E., and Oswald, A. J. 1994. Unhappiness and Unemployment. *The Economic Journal* 104 (May): 648–659. dx.doi.org/10.2307/2234639

DeHon, A.; Karel, B.; Knight Jr, T. F.; Malecha, G.; Montagu, B.; Morisset, R.; Morrisett, G.; Pierce, B. C.; Pollack, R.; Ray, S.; Shivers, O.; and Smith, J. M. 2011. Preliminary Design of the SAFE Platform. In *PLOS '11: Proceedings of the 6th Workshop on Programming Languages and Operating Systems.* New York: Association for Computing Machinery. dx.doi.org/10.1145/2039239.2039245

Dennis, L. A.; Fisher, M.; Lincoln, N. K.; Lisitsa, A.; and Veres, S. M. 2013. Practical Verification of Decision-Making in Agent-Based Autonomous Systems. ArXiv Preprint ArXiv:1310.2431. Ithaca, NY: Cornell University Library.

Docherty, B. L. 2012. *Losing Humanity: The Case Against Killer Robots.* New York: Human Rights Watch.

Fallenstein, B., and Soares, N. 2014. Vingean Reflection: Reliable Reasoning for Self-Modifying Agents. Technical Report, Machine Intelligence Research Institute, Berkeley, CA.

Fisher, K. 2012. HACMS: High Assurance Cyber Military Systems. In *Proceedings of the 2012 ACM Conference on High Integrity Language Technology,* 51–52. New York: Association for Computing Machinery. dx.doi.org/10.1145/2402676.2402695

Frey, C., and Osborne, M. 2013. The Future of Employment: How Susceptible Are Jobs to Computerisation? Technical Report, Oxford Martin School, University of Oxford, Oxford, UK.

Glaeser, E. L. 2014. Secular Joblessness. In *Secular Stagnation: Facts, Causes, and Cures,* ed. C. Teulings and R. Baldwin, 69–82. London: Centre for Economic Policy Research (CEPR)

Good, I. J. 1965. Speculations Concerning the First Ultraintelligent Machine. *Advances In Computers* 6(1965): 31–88. dx.doi.org/10.1016/S0065-2458(08)60418-0

Halpern, J. Y., and Pass, R. 2011. I Don't Want to Think About It Now: Decision Theory with Costly Computation. ArXiv Preprint ArXiv:1106.2657. Ithaca, NY: Cornell University Library. dx.doi.org/10.1111/tops.12088

Halpern, J. Y., and Pass, R. 2013. Game Theory with Translucent Players. ArXiv Preprint ArXiv:1308.3778. Ithaca, NY: Cornell University Library.

Halpern, J. Y.; Pass, R.; and Seeman, L. 2014. Decision Theory with Resource-Bounded Agents. *Topics In Cognitive Science* 6(2): 245–257.

Hetschko, C.; Knabe, A.; and Schöb, R. 2014. Changing Identity: Retiring from Unemployment. *The Economic Journal* 124(575): 149–166. dx.doi.org/10.1111/ecoj.12046

Hexmoor, H.; McLaughlan, B.; and Tuli, G. 2009. Natural Human Role in Supervising Complex Control Systems. *Journal of Experimental & Theoretical Artificial Intelligence* 21(1): 59–77. dx.doi.org/10.1080/09528130802386093

Hibbard, B. 2012. Avoiding Unintended AI Behaviors. In *Artificial General Intelligence,* Lecture Notes in Artificial Intelligence volume 7716, ed. J. Bach, B. Goertzel, and M. Iklé, 107–116. Berlin: Springer. dx.doi.org/10.1007/978-3-642-35506-6_12

Hibbard, B. 2014. Ethical Artificial Intelligence. ArXiv Preprint ArXiv: arXiv:1411.1373. Ithaca, NY: Cornell University Library.

Hibbard, B. 2015. Self-Modeling Agents and Reward Generator Corruption. In *Artificial Intelligence and Ethics: Papers from the AAAI 2015 Workshop,* ed. T. Walsh, 61–64, AAAI Technical Report WS-15-02. Palo Alto, CA: AAAI Press.

Hintze, D. 2014. Problem Class Dominance in Predictive Dilemmas. Honors Thesis, Barrett, the Honors College, Arizona State University, Tempe, AZ.

Horvitz, E. 2014. One-Hundred Year Study of Artificial Intelligence: Reflections and Framing, White paper, Stanford University, Stanford, CA (ai100.stanford.edu).

Horvitz, E., and Selman, B. 2009. Interim Report from the Panel Chairs: AAAI Presidential Panel on Long Term AI Futures. AAAI Panel held 21–22 February, Pacific Grove, CA. (www.aaai.org/Organization/Panel/panel-note.pdf)

Horvitz, E. J. 1987. Reasoning About Beliefs and Actions Under Computational Resource Constraints. Paper presented at the Third Workshop on Uncertainty in Artificial Intelligence, Seattle, WA, July 12.

Klein, G.; Elphinstone, K.; Heiser, G.; Andronick, J.; Cock, D.; Derrin, P.; Elkaduwe, D.; Engelhardt, K.; Kolanski, R.; Norrish, M.; Sewell, T.; Tuch, H.; and Winwood, S. 2009. SeL4: Formal Verification of an OS Kernel. In *Proceedings of the 22nd ACM SIGOPS Symposium on Operating Systems Principles,* 207–220. New York: Association for Computing Machinery. dx.doi.org/10.1145/1629575.1629596

Lane, T. D. 2000. Machine Learning Techniques for the Computer Security Domain of Anomaly Detection. Ph.D. Dissertation, Department of Electrical Engineering, Purdue University, Lafayette, IN.

LaVictoire, P.; Fallenstein, B.; Yudkowsky, E.; Barasz, M.; Christiano, P.; and Herreshoff, M. 2014. Program Equilibrium in the Prisoner's Dilemma Via Löb's Theorem. In Multiagent Interaction Without Prior Coordination: Papers from the 2014 AAAI Workshop. Technical Report WS-14-09. Palo Alto, CA: AAAI Press.

Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C.; and Byers, A. H. 2011. Big Data: The Next Frontier for Innovation, Competition, and Productivity. Report (May). Washington, D.C.: McKinsey Global Institute.

Manyika, J.; Chui, M.; Bughin, J.; Dobbs, R.; Bisson, P.; and Marrs, A. 2013. Disruptive Technologies: Advances That Will Transform Life, Business, and the Global Economy. Report (May). Washington, D.C.: McKinsey Global Institute.

Mokyr, J. 2014. Secular Stagnation? Not in Your Life. In *Secular Stagnation: Facts, Causes and Cures,* ed. C. Teulings and R. Baldwin, 83. London: Centre for Economic Policy Research (CEPR)

Ng, A. Y., and Russell, S. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the 17th International Conference on Machine Learning,* 663–670. San Francisco: Morgan Kaufmann

Nilsson, N. J. 1984. Artificial Intelligence, Employment, and Income. *AI Magazine* 5(2): 5.

Omohundro, S. M. 2007. The Nature of Self-Improving Artificial Intelligence. Talk presented at the Singularity Summit, San Francisco, CA 8–9.

Orseau, L., and Ring, M. 2012. Space-Time Embedded Intelligence. In *Artificial General Intelligence, 5th International Conference* (AGI 2012), ed. J. Bach, B. Goertzel, Matthew Iklé, 209–218. Berlin: Springer. dx.doi.org/10.1007/978-3-642-35506-6_22

Parasuraman, R.; Sheridan, T. B.; and Wickens, C. D. 2000. A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans.* 30(3): 286–297. dx.doi.org/10.1109/3468.844354

Platzer, A. 2010. *Logical Analysis of Hybrid Systems: Proving Theorems for Complex Dynamics,* Lecture Notes in Computer Science Volume 7386. Berlin: Springer.

Press, A. 1933. Atom-Powered World Absurd, Scientists Told. *New York Herald Tribune* September 12, p. 1.

Probst, M. J., and Kasera, S. K. 2007. Statistical Trust Establishment in Wireless Sensor Networks. In *Proceedings of the 2007 IEEE International Conference on Parallel and Distributed Systems,* volume 2, 1–8. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/ICPADS.2007.4447736

Pulina, L., and Tacchella, A. 2010. An Abstraction-Refinement Approach to Verification of Artificial Neural Networks. In *Computer Aided Verification,* Lecture Notes in Computer Science Volume 6174, 243–257. Berlin: Springer. dx.doi.org/10.1007/978-3-642-14295-6_24

Reuters. 1956. Space Travel 'Utter Bilge.' *The Ottawa Citizen*, January 3, p. 1.

Rieck, K.; Trinius, P.; Willems, C.; and Holz, T. 2011. Automatic Analysis of Malware Behavior Using Machine Learning. *Journal of Computer Security* 19(4): 639–668.

Roff, H. M. 2013. Responsibility, Liability, and Lethal Autonomous Robots. In *Rout-*

ledge Handbook of Ethics and War: Just War Theory in the 21st Century, 352. New York: Routledge Taylor and Francis Group.

Roff, H. M. 2014. The Strategic Robot Problem: Lethal Autonomous Weapons in War. *Journal of Military Ethics* 13(3): 211–227. dx.doi.org/10.1080/15027570.2014.975010

Russell, S. 1998. Learning Agents for Uncertain Environments. In *Proceedings of the Eleventh Annual Conference On Computational Learning Theory,* 101–103. New York: Association for Computing Machinery. dx.doi.org/10.1145/279943.279964

Russell, S., and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach* Third Edition. New York: Pearson, Inc.

Russell, S. J., and Subramanian, D. 1995. Provably Bounded-Optimal Agents. *Journal of Artificial Intelligence Research* 2: 575–609.

Sabater, J., and Sierra, C. 2005. Review on Computational Trust and Reputation Models. *Artificial Intelligence Review* 24(1): 33–60. dx.doi.org/10.1007/s10462-004-0041-5

Schumann, J. M., and Liu, Y. 2010. *Applications of Neural Networks in High Assurance Systems,* Studies in Computational Intelligence Volume 268. Berlin: Springer. dx.doi.org/10.1007/978-3-642-10690-3

Shanahan, M. 2015. *The Technological Singularity*. Cambridge, MA: The MIT Press.

Singer, P. W., and Friedman, A. 2014. *Cybersecurity: What Everyone Needs to Know.* New York: Oxford University Press.

Soares, N. 2014a. Formalizing Two Problems of Realistic World-Models, Technical Report, Machine Intelligence Research Institute, Berkeley, CA.

Soares, N. 2014b. The Value Learning Problem, Technical Report, Machine Intelligence Research Institute, Berkeley, CA.

Soares, N., and Fallenstein, B. 2014a. Aligning Superintelligence with Human Interests: A Technical Research Agenda, Technical Report, Machine Intelligence Research Institute, Berkeley, CA.

Soares, N., and Fallenstein, B. 2014b. Questions of Reasoning Under Logical Uncertainty, Technical Report, Machine Intelligence Research Institute, Berkeley, CA. (intelligence.org/files/QuestionsLogical Uncertainty.pdf).

Soares, N., and Fallenstein, B. 2014c. Toward Idealized Decision Theory, Technical Report, Machine Intelligence Research Institute, Berkeley, CA.

Soares, N.; Fallenstein, B.; Yudkowsky, E.; and Armstrong, S. 2015. Corrigibility. In *Artificial Intelligence and Ethics,* ed. T. Walsh, AAAI Technical Report WS-15-02. Palo Alto, CA: AAAI Press.

Spears, D. F. 2006. Assuring the Behavior of Adaptive Agents. In *Agent Technology from a Formal Perspective,* NASA Monographs in Systems and Software Engineering, ed. C. Rouff, M. Hinchey, J. Rash, W. Truszkowski, D. Gordon-Spears, 227–257. Berlin: Springer. dx.doi.org/10.1007/1-84628-271-3_8

Sullins, J. P. 2011. Introduction: Open Questions in Roboethics. *Philosophy & Technology* 24(3): 233–238. dx.doi.org/10.1007/s13347-011-0043-6

Taylor, B. J. E. 2006. *Methods and Procedures for the Verification and Validation of Artificial Neural Networks*. Berlin: Springer.

Tegmark, M. 2015. Friendly Artificial Intelligence: the Physics Challenge. In *Artificial Intelligence and Ethics,* ed. T. Walsh, AAAI Technical Report WS-15-02, 87–89. Palo Alto, CA: AAAI Press.

Tennenholtz, M. 2004. Program Equilibrium. *Games and Economic Behavior* 49(2): 363–373. dx.doi.org/10.1016/j.geb.2004.02.002

UNIDIR. 2014. The Weaponization of Increasingly Autonomous Technologies: Implications for Security and Arms Control. UNIDIR Report No. 2. Geneva, Switzerland: United National Institute for Disarmanent Research.

Van Parijs, P. 1992. *Arguing for Basic Income. Ethical Foundations for a Radical Reform.* New York: Verso.

Vinge, V. 1993. The Coming Technological Singularity. In VISION-21 Symposium, NASA Lewis Research Center and the Ohio Aerospace Institute. NASA Technical Report CP-10129. Washington, DC: National Aeronautics and Space Administration.

Vladeck, D. C. 2014. Machines Without Principals: Liability Rules and Artificial Intelligence. *Washington Law Review* 89(1): 117.

Weaver, N. 2013. Paradoxes of Rational Agency and Formal Systems That Verify Their Own Soundness. ArXiv Preprint ArXiv:1312.3626. Ithaca, NY: Cornell University Library.

Weld, D., and Etzioni, O. 1994. The First Law of Robotics (A Call to Arms). In *Proceedings of the Twelfth National Conference on Artificial Intelligence,* 1042–1047. Menlo Park, CA: AAAI Press.

Widerquist, K.; Noguera, J. A.; Vanderborght, Y.; and De Wispelaere, J. 2013. *Basic Income: An Anthology of Contemporary Research.* New York: Wiley/Blackwell.

Winfield, A. F.; Blum, C.; and Liu, W. 2014. Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection. In *Advances in Autonomous Robotics Systems, 15th Annual Conference*. Lecture Notes in Artificial Intelligence, ed. M. Mistry, A. Leonardis, M. Witkowski, and C. Melhuish, 85–96. Berlin: Springer.

Wissner-Gross, A., and Freer, C. 2013. Causal Entropic Forces. *Physical Review Letters* 110(16): 168702. dx.doi.org/10.1103/PhysRevLett.110.168702

Yampolskiy, R. 2012. Leakproofing the Singularity: Artificial Intelligence Confinement Problem. *Journal of Consciousness Studies* 19(1–2): 1–2.

Stuart Russell is a professor of computer science at the University of California, Berkeley. His research covers many aspects of artificial intelligence and machine learning. He is a fellow of AAAI, ACM, and AAAS and winner of the IJCAI Computers and Thought Award. He held the Chaire Blaise Pascal in Paris from 2012 to 2014. His book *Artificial Intelligence: A Modern Approach* (with Peter Norvig) is the standard text in the field.

Daniel Dewey is the Alexander Tamas Research Fellow on Machine Superintelligence and the Future of AI at Oxford's Future of Humanity Institute, Oxford Martin School. He was previously at Google, Intel Labs Pittsburgh, and Carnegie Mellon University.

Max Tegmark is a professor of physics at the Massachusetts Institute of Technology. His current research is at the interface of physics and artificial intelligence, using physics-based techniques to explore connections between information processing in biological and engineered systems. He is the president of the Future of Life Institute, which supports research advancing robust and beneficial artificial intelligence.