

AI RFI Responses, October 26, 2018

Update to the 2016 National Artificial Intelligence Research and Development Strategic Plan RFI Responses

DISCLAIMER: The [RFI public responses](#) received and posted do not represent the views and/or opinions of the U.S. Government, National Science and Technology Council (NSTC) Select Committee on Artificial Intelligence (AI), NSTC Subcommittee on Machine Learning and AI, NSTC Subcommittee on Networking and Information Technology Research and Development (NITRD), NITRD National Coordination Office, and/or any other Federal agencies and/or government entities. We bear no responsibility for the accuracy, legality or content of all external links included in this document.

RFI Response: National Artificial Intelligence Research and Development Strategic Plan

Institute for Defense Analyses Systems and Analyses Center (IDA SAC)

Executive Summary and Introduction: Proposed Changes

Page 3, Executive Summary, paragraph 1. Replace with:

“Artificial intelligence (AI) is the automation of tasks that have historically required human intelligence. AI is enabled by a broad range of transformative technologies (AI-enabling technologies) whose applications hold promise for tremendous societal and economic benefit. AI has the potential to revolutionize how we live, work, learn, discover, and communicate. AI can further our national priorities, including increased economic prosperity, improved educational opportunities and quality of life, and enhanced national and homeland security. Because of these potential benefits, the U.S. government has invested in AI research for many years. Yet, as with any significant field in which the Federal government has interest, there are not only tremendous opportunities but also a number of considerations that must be taken into account in guiding the overall direction of federally funded R&D in AI.”

Rationale: *Artificial Intelligence* is a term that is used in different ways – to refer to an academic field of study, to refer to systems that are intended to exhibit some level of “intelligence,” and sometimes to refer to specific technologies that enable the intelligent capabilities of AI systems. Research in AI includes research in technologies as diverse as those for creating neuromorphic chips and those used in deep learning. It is not appropriate to refer to AI-enabling technologies as a singular technology.

Conforming changes:

Paragraph 3: Replace “AI knowledge and technologies” with “discoveries and AI-enabling technologies”.

Throughout: Replace “AI technologies” with “AI-enabling technologies” to emphasize AI as a goal, not as a technology or set of technologies.

Page 3, Strategy 3. Replace with:

Strategy 3: Take a leadership role in understanding and addressing the ethical, legal, and societal implications of AI. The acceptable uses of AI will be informed by the tenets of law and ethics; the challenge is determining how to apply those tenets to these new technologies, particularly those involving autonomy, agency, and control. To meet this challenge, the United States should lead a coalition of the world’s liberal democracies in understanding, creating, and using AI within Western legal and ethical parameters.

Page 3, Strategy 7. Replace with:

Strategy 7: Build the Needed National AI R&D Workforce. Attaining the AI R&D advances outlined in this strategy requires a national AI R&D workforce comprised of highly educated AI researchers, well-informed program managers, and well-trained developers. The accelerated pace of change associated with AI is straining the education and workforce systems' capacity to educate, train, and hire individuals with the appropriate expertise and knowledge.

Page 14, last paragraph. Replace with:

“AI research may be at the beginning of a possible third wave, which will focus on explanation and extrapolation. The goals of this research are to allow humans to interact with learned models through an explanation and correction interface, to clarify the basis for and reliability of outputs, to establish appropriate levels of trust in AI systems, and to broaden current narrow AI capabilities to ones that can generalize across broader task domains. If successful, engineers could create systems that construct explanatory models for classes of real world phenomena, communicate with people in natural ways, learn and reason as they encounter new tasks and situations, and solve novel problems (with accompanying explanations) by generalizing from past experience.”

Rationale: The original text stated that the goal of the 3rd wave included moving beyond narrow AI. However, the original document also acknowledges that most AI researchers believe that truly General AI is still decades away. Moving beyond narrow AI indicates moving to General AI in the context of this dichotomy. Hence this goal as stated sounds too ambitious for this third wave, where we do not expect to achieve truly General AI. Thus, we suggest rewording it to refer to broadening current narrow AI capabilities.

Strategy 2 Proposed Changes

Page 23, starting with the last sentence of the second paragraph: “To address these concerns....support humans during periods of excessive workload or fatigue.”

Suggested change: Delete sentences in this range, including the four general principles.

Rationale: These sentences and principles address the research on human-automation interaction in general. The four issues discussed – interface design, displays, automation flexibility, and operator training – are relevant to situations where any technology is coupled with a human. All the references cited are from research on human-automation interaction and not specifically on AI.

Page 24, “Developing techniques for visualization and AI-human interfaces.”

Suggested change: Delete entire section.

Rationale: This is a general paragraph on why good interface design and visualization are important when humans interact with technology, followed by some examples of scenarios in

different domains. There is no connection made between this section on interface design and AI. Although visualizations can be important to interfaces between humans and AI systems, no visualization issues that would be specific to AI research and development are raised.

Page 25, “Developing more effective language processing systems.”

Suggested change: The purpose of the DARPA/Siri example is not clear. If the purpose is to illustrate public benefits of past government research, then move the example to the general Introduction. If the section is intended to be a tutorial on the importance of AI language processing, move it to “Current State of AI” section and revisit it briefly in this section.

Strategy 3 Proposed Changes

Page 26: The proposed revised title and body appears below, in lieu of line-in/line-out.

Strategy 3: Take a leadership role in understanding and addressing the ethical, legal, and societal implications of AI

The acceptable uses of AI will be informed by the tenets of law and ethics; the challenge is how to apply those tenets to these new technologies, particularly those involving autonomy, agency, and control. To meet this challenge, the United States should lead a coalition of the world’s liberal democracies in understanding, creating, and using AI within Western legal and ethical parameters. This means promoting multilateral respect, maintenance, and enhancement of the social contract – the core of Western democracy.

By enunciating and defining the legal boundaries within which AI might be employed, the coalition will inform the design and behavior of AI systems. Policy makers, ethicists, and judges regularly scrutinize the acceptable use of new technologies, and AI is no different. However, the challenge is applying legal, moral, and ethical scrutiny to technologies exhibiting human behaviors of autonomy, agency, and control. The dominant research should focus on both understanding the ethical, legal, and social implications of AI, as well as aligning the usage and employment of AI to those principles underpinning the social contract. For example, the concepts of individual privacy must also be taken into account.¹

“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 dimension is tied intimately to questions of what engineering techniques are available, how reliable these techniques are, and what trade-offs are made—all areas where computer science, machine learning, and broader AI expertise is valuable.”

To achieve these ends, specific investments in research and development should be made in the form of recruiting and soliciting experts from various disciplines and industries: law,

¹ Further information on this issue can be found in the National Privacy Research Strategy.

philosophy, computer science, information technology, social and behavioral psychology, biomedicine, and representatives from the militaries of the coalition. As noted by the Future Life Institute:²

The following subsections explore key research challenges in this area requiring the cooperation of the above mentioned disciplines.

Ensuring fairness, transparency, and accountability-by-design.

There are serious theoretical and practical issues about how to represent and “encode” value and belief systems. Scientists must study to what extent justice and fairness considerations can be designed into AI systems and how to accomplish this within the bounds of engineering techniques. Many concerns have been voiced about the susceptibility of AI machine learning algorithms to error and misuse; misunderstanding by the human beings who interface with them; and the possible ramifications for discrimination based on gender, age, racial, or economic classes. The proper collection and use of data for AI systems represent an important challenge. Beyond purely data-related issues, however, larger questions arise about the design of AI systems to be inherently just, fair, transparent, and accountable. Researchers must learn how to design these systems so that their actions and decision-making are transparent and easily interpretable by humans and thus can be examined for any bias they may contain, rather than just learning and repeating these biases.

Building ethical AI

Ethical issues vary according to culture, religion, and beliefs. Ethics is inherently a philosophical question, whereas AI technology depends on, and is limited by, engineering. Ethical principles are typically stated with varying degrees of vagueness and are difficult to translate into precise system and algorithm design. Therefore, researchers must strive to develop algorithms and architectures that are verifiably consistent with, or conform to, existing laws, social norms and ethics – clearly a very challenging task.

In addition to the fundamental assumptions of justice and fairness, AI needs adequate methods for values-based conflict resolution in which the system incorporates principles that can address the realities of complex situations where strict rules are impracticable. For example:

- How might advances in AI frame new “machine-relevant” questions in ethics, or what uses of AI might be considered unethical?
- How do AI systems, particularly with new kinds of autonomous decision-making algorithms, resolve moral dilemmas based on independent and possibly conflicting value systems?

Multi-disciplinary based reference frameworks can be developed to guide AI system reasoning and decision-making, in order to explain and justify its conclusions and actions. The same multi-

² “An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence,” The Future of Life Institute, <http://futureoflife.org/ai-open-letter/>.

disciplinary approach can be used to develop datasets and knowledge bases that reflect appropriate value systems, including examples that indicate preferred behavior when presented with difficult moral issues or with conflicting values.

Designing architectures for ethical AI

Researchers will need to focus on how to best address the overall design of AI systems that align with ethical, legal, and societal goals. Additional progress in fundamental research must be made to determine how to best design architectures for AI systems that incorporate ethical reasoning. A variety of approaches have been suggested, such as a two-tier monitor architecture that separates the operational AI from a monitor agent responsible for the ethical or legal assessment of any operational action.³ An alternative view is that safety engineering is preferred: A precise conceptual framework for the AI agent architecture is used to ensure that AI behavior is safe and not harmful to humans.⁴ A third method is to formulate an ethical architecture using set theoretic principles combined with logical constraints on AI system behavior that restrict action to conform to ethical doctrine.⁵ As AI systems become more general, their architectures will likely include subsystems that can take on ethical issues at multiple levels of judgment, including rapid response pattern matching rules, deliberative reasoning for slower responses for describing and justifying actions, social signaling to indicate trustworthiness for the user, and social processes that operate over even longer time scales to enable the system to abide by cultural norms.⁶

Rationale: The major change to this section is the suggestion that the US should clearly state that it is going to take a leadership position in understanding and addressing the Legal and Ethical (L&E) issues. This is important for many reasons. First, our adversaries (as described in the national security strategy) do not have the same legal or ethical constraints as we do. But our strength is in our association with our liberal democratic allies in regards to the social contract we have with our people. This should be used to our advantage in distinguishing ourselves from our adversaries. Worse case, as different countries with different ethical standards (or no standards) begin to employ AI, it may result in a race to the bottom, manifesting in illegal, unethical, and dangerous permissions given to AI. Second, from a more pragmatic standpoint, these issues should be aligned, as much as feasible, in order for us to cooperatively work with our allies on security, finance, trade, diplomacy, intelligence, and military issues. In the past, even small distinctions in our approach to laws have hampered our

³ A. Etziona and O. Etzioni, "Designing AI Systems that Obey Our Laws and Values", *Communications of the ACM* 59 (9), (2016):29-31.

⁴ R. Y. Yampolsky, "Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach," in *Philosophy and Theory of Artificial Intelligence*, edited by V.C. Muller (ed.), (Heidelberg: Springer Verlag: 2013), 389–96.

⁵ R. C. Arkin, "Governing Legal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture," Georgia Institute of Technology Technical Report, GIT-GVU-07-11, 2007.

⁶ B. Kuipers, "Human-like Morality and Ethics for Robots", AAIL-16 Workshop on AI, Ethics and Society, 2016.

sharing of valuable information. In addition, a multi-disciplinary approach must be used in a capability that impacts everything from laws to societal engagements.

“Improving Fairness” has been changed to “Ensuring Fairness.” If “fair is fair,” it may be impossible to improve fairness, but it is important to ensure fairness. In other places within the text, we brought the lead thought to the front of the paragraph so that the reader immediately understands the intent.

Strategy 4 Proposed Changes

Page 27: A proposed revised title and body appears below, in lieu of line-in/line-out.

Strategy 4: Assure the Dependability of AI Systems

Before an AI system is put into widespread use, assurance is needed that the system will operate safely and securely, in a controlled manner, and in all relevant circumstances. Research is needed to address this challenge of creating robust AI systems that are reliable, dependable, trustworthy, and safe. Even more than other complex systems, assuring dependability is a challenge for AI systems due to:⁹¹

- *Complex and uncertain environments:* In many cases, AI systems are designed to operate in complex environments with a large number of potential states that cannot be exhaustively (or even statistically) examined or tested. Systems will confront conditions that were never considered during design or evaluated during testing.
- *Unpredictable behavior:* Because the system’s response is a high-dimensional function of its inputs and state, it is very difficult to characterize how a system will respond in every possible situation. For AI systems that learn after deployment, there is even more uncertainty, as a system’s behavior may be strongly influenced by periods of learning under unsupervised conditions. Many researchers in autonomous systems consider the inherent unpredictability of system responses to be so high that the system can be thought of as nondeterministic.
- *Goal misspecification:* Due to the difficulty of translating human goals into computer instructions, the goals that are programmed for an AI system may not match the goals that were intended by the programmer. This is in part due to the fact that the AI system’s world model is much simpler than the human’s. Goal misspecification can also include the effects of training data or algorithmic bias as discussed with respect to Strategy 3.
- *Human-machine interactions:* In many cases, the performance of an AI system is substantially affected by human interactions and vice versa. In these cases, the system will need to be robust to variations in how different humans interact with it.⁹² Design of the human-machine teaming protocols will be an essential part of system design.
- *Emergent behavior:* In addition to the unpredictability of individual AI systems, collections of independent AI systems and humans may exhibit emergent behaviors in which low-level choices of independent agents lead to unexpected higher-level

phenomena of the group. Flocking behavior of birds is an emergent behavior, but so is gridlock in urban traffic.

To address these issues and others, additional investments are needed to advance general approaches to the assurance of AI dependability, which must encompass not only safety⁹³ but also physical security, cybersecurity, and robustness of performance. Effective assurance cases for AI dependability will require architectures and designs that explicitly support assurance; explainability and transparency of enabling technologies; verification of appropriate trust; and novel approaches to test, evaluation, verification, and validation of systems.

Enhancing verification and validation

New methods are needed for verification and validation of AI systems. *Verification* establishes that a system meets formal specifications, whereas *validation* establishes that a system meets the user's operational needs. Safe and dependable AI systems may require new means of assessment (determining if the system is malfunctioning, perhaps when operating outside expected parameters), diagnosis (determining the causes for the observed behavior), and recovery (enabling the system to self-correct or override undesired behavior). For systems operating autonomously over extended periods of time, system designers cannot consider every condition the system might encounter. Such systems will need to possess capabilities for self-assessment, self-diagnosis, and self-correction in order to be robust and reliable.

Furthermore, new test and evaluation techniques will need to be developed to verify and validate the adequacy of these capabilities in operation. Research into novel uses of modeling and simulation and introspective instrumentation of AI systems will be needed to accomplish this. AI systems using neural network technologies may also benefit from more work to establish theoretical foundations that could help verify that they will operate as intended.

Improving explainability, transparency, and trust

As noted under Strategy 3, transparency of AI behavior may be essential for social acceptance and legal approval of AI systems. This is closely related to the assurance challenges of AI. Many current AI algorithms, such as those based on deep learning, are opaque to users, with few existing mechanisms for explaining their results. This is especially problematic for domains such as healthcare, where doctors need explanations to justify a particular diagnosis or a course of treatment. Enabling techniques such as decision-tree induction provide a form of built-in explanation but are generally less powerful as predictors than other, less transparent techniques. To achieve assurance of system dependability and justified trust from human users and regulatory authorities, researchers must not only develop systems that are transparent and intrinsically capable of explaining the reasons for their behavior to users, they must also develop new ways to collect and combine evidence of system dependability convincingly.

[Retain existing sections "Securing against attacks" and "Achieving long-term AI safety" as is, but change the title of the latter to "Achieving long-term assured dependability of AI"]

Rationale: Safety and (cyber) security are special cases of the broader assurance challenge for AI and autonomous systems. The R&D challenge for the AI community is the broader question of how to convince regulatory bodies, insurers, and users that AI-enabled systems will behave as they ought, and not behave in unacceptable ways. Treating safety and security in stove-piped fashion, as has been the rule in the past, will not suffice. The digression about software quality and productivity is not directly related to explanation or transparency and is only indirectly related to possible issues of trust. If current software capabilities cannot support AI R&D adequately, investment in improved software processes should be addressed explicitly under Strategy 1, long-term enabling research.

Strategy 7 Proposed Changes

Page 35: A proposed revised title and body appears below, in lieu of line-in/line-out.

Strategy 7: Build the Needed National AI R&D Workforce

Attaining the AI R&D advances outlined in this strategy requires a national AI R&D workforce comprised of highly educated AI researchers, well-informed program managers, and well-trained developers. AI researchers, having earned graduate degrees in STEM and IT fields, will be in high demand in industry, government, and academia. The accelerated pace of change associated with AI (and other technical innovations) is straining the education and workforce systems' capacity to educate, train, and hire individuals with the appropriate expertise and knowledge.⁷ To compete in the international race to develop our “best and brightest” in AI, the U.S. must adapt to the rapidly changing nature of work, and invest in highly educated specialists in AI. Diversity issues should also be explored since studies have shown that a diverse workforce can lead to improved outcomes.⁸

Data is needed to characterize the current state of the AI R&D workforce in academia, government, and the private sector, and to predict the evolving supply and demand for AI talent. While no official statistics on the current and future AI workforce exist, recent reports from various commercial and academic sectors cite a current shortage of experts in AI⁹ with demand expected to continue to escalate.¹⁰ High tech companies are reportedly investing significant resources into recruiting faculty members and students with AI expertise.¹¹ Higher education and the private sector are competing to recruit and retain AI talent in various fields

⁷ Artificial Intelligence: Emerging Opportunities, Challenges, and Implications. Report of the Committee on Science, Space, and Technology, House of Representatives. March 2018. United States Government Accountability Office. GAO-18-142SP.

⁸ J.W. Woody, C.M. Beise, A.B. Woszczyński, and M.E. Myers, Diversity and the information technology workforce: Barriers and opportunities. *Journal of Computer Information Systems*, 43, (2003): 63-71.

⁹ “Startups Aim to Exploit a Deep-Learning Skills Gap”, *MIT Technology Review*, January 6, 2016.

¹⁰ AI talent grab sparks excitement and concern”, *Nature*, April 26, 2016.

¹¹ “Artificial Intelligence Experts are in High Demand”, *The Wall Street Journal*, May 1, 2015.

(e.g., machine-learning, robotics, and natural language processing).¹² At the same time the U.S. government is predicting increased demand for highly educated STEM specialists to serve in critical roles in AI and other fields such as cybersecurity, data engineering, and quantum information science. U.S. citizens with AI expertise will be required for R&D in AI applications addressing national security concerns.

Mechanisms are needed to ensure an adequate supply of AI workforce talent to serve in industry, academia, and government roles in the United States. The pipeline that prepares members of this future AI workforce begins in K-12 education. The U.S. needs to invest significantly in its schools, teachers, and students to motivate, challenge, mentor, and support diverse learners in STEM disciplines, in school as well as after-school and summer programs, internships, and in funded higher education. In response to the narrowing of the pipeline of STEM scholars that occurs from elementary school to high school, and into undergraduate and graduate education, concerted efforts are needed to identify and nurture talent and interest in STEM fields early and consistently. To increase the enrollment of qualified undergraduate and graduate students in critical disciplines in science, technology, computing, engineering, and mathematics, and their completion of degree programs, academic programs must transform to address factors identified in a 2018 National Academy of Science report associated with negative culture, incentives, and practices of graduate education in STEM fields. To prepare the needed AI workforce, higher education must identify and nurture interested and talented STEM scholars who are diverse in gender, race and ethnicity, nation of origin, disability, and socioeconomic background.¹³

Rationale: The strategy title is revised to reflect the need for analysis of and for investment in the AI R&D workforce. There is a need to build the AI workforce, not just to analyze it. The revised text integrates several of the original strategy's main points, sentences, and references about the demand for AI R&D personnel, and the data to support the demand. In response to the demand for highly educated and specialized AI workforce talent, it is imperative to grow AI (and other STEM) talent in U.S. K-12 schools and higher education. This imperative includes U.S. citizens to research, design, and develop national security AI applications.

Because graduate education is a significant player in developing the workforce, reference is made to the 2018 National Academy of Science report that recommends significant structural changes in graduate education to meet the nation's STEM education and workforce needs.

¹² "Million dollar babies: As Silicon Valley fights for talent, universities struggle to hold on to their stars", *The Economist*, April 2, 2016.

¹³ *Graduate STEM Education for the 21st Century*. Alan Leshner and Layne Scherer, Editors. The National Academic Press. ISBN 978-0-309-47273-9/DOI 10.17226/25038

Proposed Additional Recommendation:

Identify the R&D areas and projects of particular importance to the nation that are unlikely to be addressed by commercial or academic efforts.

Progress in the commercial and academic sectors will be an important, perhaps crucial, element in the nation's progress at expanding both the set and the scope of AI-enabling technologies. Nevertheless, there are potential public benefits of AI that neither the commercial nor academic sectors will be incentivized to pursue, and others that will not be feasible without enabling R&D underwritten by government. The Federal government should therefore emphasize AI investments in areas of strong societal importance that are not aimed at consumer markets—areas such as AI for public health, urban systems and smart communities, social welfare, criminal justice, environmental sustainability, and national security, as well as long-term research that accelerates the production of AI capabilities and underlying technologies.

Note: Superscripted numbers without footnotes are references to the footnotes in the original 2016 strategy.