International Journal of Advanced Technology in Engineering and Science
Vol. No.3, Issue No. 12, December 2015
www.ijates.com
ijates
ISSN 2348 - 7550

ARTIFICIAL INTELLIGENCE

Heena Budhiraja[1], Gaurav[2], Mannu[3]

[1,2,3]IT, Hindu College of Engineering, Sonepat , Haryana (India))

ABSTRACT

This paper will review certain approaches to artificial intelligence work, mainly work done since 1960.An important area of research involves designing a machine that can adequately improve its own performance, as well as solve other problems normally requires human intelligence. There is, in some quarters, concern about high–level machine

Intelligence and super intelligent AI coming up in a few decades, bringing with it significant risks for humanity. In other quarters, these issues are ignored or considered science fiction. We wanted to clarify what the distribution of opinions actually is, what probability the best experts currently assign to high–level machine intelligence coming up within a

Particular time–frame, which risks they see with that development, and how fast they see these developing. The median estimate of respondents was for a one in two chance that high level machine intelligence will be developed around 2040-2050, rising to a nine in ten chance by 2075. Experts expect that systems will move on to super intelligence in less than 30 years thereafter. They estimate the chance is about one in three that this development turns out to be 'bad' or 'extremely bad' for humanity.

I. INTRODUCTION

Artificial Intelligence is the study of how to make computer to do things which at a moment people do better.

We can divide artificial intelligence into 4 main categories:

1. System that thinks like humans-it is the automation of activities that we associate with human thinking. It includes activities such as Decision making, problem solving and learning.

There are two ways to determine how human thinks. The first is through Interroception and second is through cytological experiments.

2. System that think rationally-It means right thinking. In this we design a system which draws correct conclusions from correct premises.

3. System that acts like human-It is the study of how to make computers to do things which at the moment people do better.

4. Systems that act rationally-It is the study of design of intelligence agents. An intelligent agent is one that acts so as to achieve the best outcome.

Artificial Intelligence began with the " conjecture that every aspect of learning or any other feature of intelligence is the principle to be so precisely described that a machine can be made to simulate it." and moved swiftly from this vision to grand promises for general human-level AI within a few decades. This vision of general AI has now become merely a long-term guiding idea for most current AI research, which focuses on specific scientific and engineering problems and maintains a distance to the cognitive sciences. This vision of

general AI has now become merely a long-term guiding idea for most current AI research, which focuses on specific scientific and engineering problems and maintains a distance to the cognitive sciences. A small minority believe the moment has come to pursue general AI directly as a technical aim with the traditional methods – these typically use the label 'artificial general intelligence'. If general AI were to be achieved, this might also lead to super intelligence: "We can tentatively define a super intelligence as *any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest*." .One idea how super intelligence might come about is that if we humans could create artificial general intelligent ability at a roughly human level, then this creation could, in turn, create yet higher intelligence, which could, in turn, create yet higher intelligence, and so on So we might generate a growth well beyond human ability and perhaps even an accelerating rate of growth: an 'intelligence explosion.[5]

II. A KNOWLEDGE-BASED APPROACH TO NATURAL LANGUAGE UNDERSTANDING

A significant feature of any natural language is that it can serve as its own meta-language. One can use a natural language to talk about the language itself as well as to give instruction in the use and understanding of the language.Because human beings are able to use their natural language to talk about that natural language itself;we have been investigating methods of knowledge representation and natural language understanding that would enable an AI system to do likewise. We have implemented a language-understanding system in the role of an educable cognitive agent whose task domain includes language understanding and whose discourse domain includes knowledge of its own language.

This system has just one (initially primitive) language, which becomes increasingly more sophisticated as the system accepts instruction expressed in its evolving language. Such a system must start with some language facility, and we have strived to make this initial kernel language as small and as independent of theory as possible. With an unbiased kernel language, teacher-users should ideally be able to bootstrap into the language of their choice.

III. LOGICAL FOUNDATIONS FOR BELIEF REPRESENTATION

Our research consists of the design and implementation of a logically and psychologically adequate computer system capable of representing and reasoning about the cognitive attitudes of intelligent agents. The agents include users,

other AI systems, and the system itself; the cognitive attitudes include beliefs, knowledge, goals, and desires. The system will be able to represent nested attitudes; it will be sensitive to the intensionality and indexicality of attitudes,

in particular, to the phenomenon of quasi-indexicality, a feature at the core of self-referential beliefs; and it will be able to expand and refine its beliefs by interacting with users in ordinary conversational situations. The system is being implemented in the SNePS Semantic Network Processing System (Shapiro, 1979) using an augmented transition network grammar for parsing and generation.

IV. ARTIFCIAL INTELLIGENCE TODAY

Artifcial 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 _elds. 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 classi_cation, 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 potentialbene_ts are huge, since everything that civilization has to o_er is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magni_ed 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 bene_ts 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.

V. SHORT-TERM RESEARCH PRIORITIES

5.1 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. 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]

1. Labor market forecasting: When and in what order should we expect various jobs to become Automated. 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 . However, increased automation may push income distribution urther towards a power law, 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.

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

3. Policy for managing adverse effects: What policies could help increasingly automated societies flourish? For example, Brynjolfiesson and McAfee 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 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 ourish? Unemployment is not the same as leisure, and there are deep links between unemployment and unhappiness, self-doubt, and isolation 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 .

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. Research on improved metrics could be useful for decision-making.

5.2 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, public policy, professional ethics, and philosophical ethics, and will require expertise from computer scientists, legal experts, political scientists, and ethicists. For example:

1. Liability and law for autonomous vehicles: If self-driving cars cut the roughly 40,000 annual UScfatalities 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 selfdriving cars best be realized Should legal questions about AI be handled by existing (softwareand internet-focused) \cyberlaw", or should they be treated separately [14]? In both military andcommercial 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 .

2. Machine ethics: How should an autonomous vehicle trade 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 be the subject of national standards.

3. Autonomous weapons: Can lethal autonomous weapons be made to comply with humanitarian law If, as some organizations have suggested, autonomous weapons should be banned is it possible to develop a precise de_nition 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 done Are

autonomousweapons likely to reduce political aversion to conict, or perhaps result in \accidental" battles or wars 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 cybersecurity and 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 .

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 the EPSRC Principles of Robotics and recently announced programs such as Stanford's One-Hundred Year Study of AI and the AAAI committee on AI impact and ethical issues.

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.

5.3 Computer Science Research for Robust AI [1]

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" [91]. Di_erent ways in which an AI system may fail to perform as desired correspond to di_erent areas of robustness research:

1. Veri_cation: how to prove that a system satis_es 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.

5.3.1 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 , a complete, general-purpose operating-system kernel that has been mathematically checked against a formal speci_cation 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 Not only should it be possible to build AI systems on top of veri_ed 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[5]. This mirrors the agent architectures used in Russell and Norvig [69], 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 veri_able agents, but we lack the formal \algebra" to properly de_ne, explore, and rankthe space of designs. Perhaps the most salient di_erence between veri_cation of traditional software and veri_cation of AI systems is that the correctness of traditional software is de_ned with respect to a _xed 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 . A lack of design-time knowledge also motivates the use of learning algorithms within the agent software, and veri_cation becomes more di_cult: statistical learning theory gives so-called _-_ (probably approximately correct) bounds, mostly for the somewhat unrealistic settings of supervised learning from 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 couldd improve our abilities in this area, possible extending work on Bayesian quadrature . Work in adaptive control theory, the theory of so-called cyberphysical systems, and veri_cation of hybrid or robotic systems [2, 93] 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 and the notion of partial programs [4, 80] allows thedesigner 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.

5.3.2 Validity

A veri_cation theorem for an agent design has the form, \If environment satis_es assumptions _ then behavior satis_es requirements ." There are two ways in which a veri_ed agent can, nonetheless, fail to be a bene_cial agent in actuality: first, the environmental assumption _ is false in the real world, leading to behavior that violates the requirements ; second, the system may satisfy the formal requirement but still behave in ways that we highly undesirable in practice. It may be the case that this undesirability is a consequence of satisfying when _ is violated; i.e., had _ held the undesirability would not have been manifested; or it may be the case that the requirement is erroneous in itself. Russell and Norvig 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 thesame dirt. [4]

The requirement should focus not on dirt cleaned up but on cleanliness of the oor. Such specifiation 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 speci_cation: the notions of \beneficial" and \desirable" are not separately made formal, so one cannot straightforwardly prove that satisfying necessarily leads to desirable behavior and a bene_cial 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-o_s can be made { all areas where computer science, machine learning, and broader AI expertise is valuable.

5.3.3 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[6].

Robustness against exploitation at the low level is closely tied to veri_ability and freedom from bugs. For example, the DARPA SAFE program aims to build an integrated hardware-software system with a exitible metadata rule engine, on which can be built memory safety, fault isolation, and other protocols that could improve security by preventing exploitable aws [20]. Such programs cannot eliminate all security laws (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, analyzing malware or detecting potential exploits in other programs through code analysis. 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-behavioraltrust establishment and computational reputation models.

5.3.4 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, 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 .

Automated vehicles are a test-bed for executive 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 humancomputer teams, both for identifying situations where control should be transferred, and for applying human judgment e_ciently to the highest-value decisions.

VI. 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 e_orts will succeed in the foreseeable future, then additional current research beyond that mentioned in the previous sections will be

motivated as exempli_ed below, to help ensure that the resulting AI will be robust and bene_cial.Assessments of this success probability vary widely between researchers, but few would argue withgreat con_dence 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", and Astronomer Royal Richard Woolley called interplanetary travel \utter bilge" in 1956. 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.

6.1 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, veri_able safety properties will become increasingly valuable. If the theory of extending veri_able 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 veri_cation 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. Attempting to straightforwardly apply formal veri_cation tools to this more general setting presents new difficulties,including the challenge that a formal system that is su_ciently powerful cannot use formal methods inthe obvious way to gain assurance about the accuracy of functionally similar formal systems, on pain ofinconsistency via Godel's incompleteness It is not yet clear whether or how this problem can be

overcome, or whether similar problems will arise with other veri_cation methods of similar strength.[3]

Finally, it is often di_cult to actually apply formal veri_cation techniques to physical systems, specially systems that have not been designed with veri_cation in mind. This motivates research pursuing a general theory that links functional speci_cation to physical states of a_airs. 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 satis_cing 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.

6.2 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 . 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 signi_cant 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, how to take into account correlations between AI systems' behaviors and those of their environments or of other agents, how agents that are embedded in their environments should reasons, and how to reason about uncertainty over logical consequences of beliefs or other deterministic computations These topics may benefit from being considered together, since they appear deeply linked .

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 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" machine ethics may not be practical, making \aligning" the values of powerful AI systems with our own values and preferences difficulty. 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 exible, 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. 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 over a viable approach, in which a system infers the preferences of another actor, assumed to be a reinforcement learner itself. Other approaches could use di_erent 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 diffcult methods could become viable, suggesting that research on such methods could be useful; for example, Bostrom reviews preliminary work on a variety of methods for specifying goals indirectly.

6.3 Security [2]

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 cyberattacksmay be extremely e_ective, while on the other hand, the use of AI and machine learning techniques along with signi_cant progress in low-level system reliability may render hardened systems much less vulnerable than today's. From a cryptographic perspective, it appears that this conict favors defenders over attackers; this may be a reason to pursue e_ective 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 [2]. Both theoreticaand practical sides of this question warrant investigation. If the general case of AI containment turns out to be prohibitively di_cult, then it may be that designing an AI system and a container in parallel imore successful, allowing the weaknesses and strengths of the design to inform the containment strategy. 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 pro_table questions in the eld of computer security.

6.4 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 diffculty of maintaining meaningful human control. 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 (and conversely, seeking unconstrained situations is sometimes a useful heuristic . This could become problematic, however, if we wish to repurpose the system, to deactivate it, or to signi_cantly alter its decision-making process; such a system would rationally avoid these changes. Systems t hat do not exhibit these behaviors have been termed corrigible systems , 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 [and theoretical frameworks could be developed to better understand the space of potential systems that avoid undesirable behaviors .

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. Hammond 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 settled 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 , 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 terrm.

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 de_ne \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 .

Stanford's One-Hundred Year Study of Artiffcial Intelligence includes \Loss of Control of AI systems" as an area of study, specifcally 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" 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 , and could extend or critique existing approaches begun by groups such as the Machine Intelligence Research Institute .

VII. CONCLUSION

In summary, success in the quest for arti_cial intelligence has the potential to bring unprecedented bene_ts to humanity, and it is therefore worthwhile to research how to maximize these bene_ts 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 bene_cial, and aligned with human interests.

REFRENCES

[1] Shoshana L.HARDT and William J.Rapaport "Recent And Cureent Artificial Intelligence, Research in the Department of Computer Science"1986"Pg 1-10".
[2] Eliezer Yudkowsky "Artificial Intelligence as a positive and negative factor in global risk " 2008"Pg 1-46".
[3] Stuart J.Russenn and Peter Norvig"Artificial Intelligence A Modern Approach "1995"Pg 1-946.
[4] Nick Bostrom and Eliezer Yudkowsky"The Ethics Of Artificial Intelligence " 2011"Pg 1-20.
[5] John MC.Carthy and Patrick J.Haygs"Some Philosophical Problem from the stand point of Artificial Intelligence " 1969" Pg 1-59.
[6] Research priorities For Robust and Beneficial Artificial Intelligence"2015"Pg1-12