AI Dangers: Imagined and Real

Devdatt Dubashi and Shalom Lappin
Chalmers University of Technology and University of Gothenburg

GoCAS Seminar
Chalmers University of Technology

May 11, 2016

Outline

The Singularity

- Good (1965) proposed that computational machines will improve in competence at an exponential rate.
- They will reach the point of the *singularity*, where they correct their own defects and program themselves to produce superintelligent artificial agents that far surpass human capacities in virtually every cognitive domain.
- Kurzweil (1999, 2011), and other futuroligists, postulate the inevitability of superintelligent artificial agents as the necessary result of the inexorable rate of progress in computational technology.
- They cite Moore's law for the exponential growth in the processing power of computer chips as the analogical basis for this claim.

The Singularity

- Good (1965) proposed that computational machines will improve in competence at an exponential rate.
- They will reach the point of the *singularity*, where they correct their own defects and program themselves to produce superintelligent artificial agents that far surpass human capacities in virtually every cognitive domain.
- Kurzweil (1999, 2011), and other futuroligists, postulate the inevitability of superintelligent artificial agents as the necessary result of the inexorable rate of progress in computational technology.
- They cite Moore's law for the exponential growth in the processing power of computer chips as the analogical basis for this claim.

The Singularity

- Good (1965) proposed that computational machines will improve in competence at an exponential rate.
- They will reach the point of the *singularity*, where they correct their own defects and program themselves to produce superintelligent artificial agents that far surpass human capacities in virtually every cognitive domain.
- Kurzweil (1999, 2011), and other futuroligists, postulate the inevitability of superintelligent artificial agents as the necessary result of the inexorable rate of progress in computational technology.
- They cite Moore's law for the exponential growth in the processing power of computer chips as the analogical basis for this claim.

The Singularity

- Good (1965) proposed that computational machines will improve in competence at an exponential rate.
- They will reach the point of the *singularity*, where they correct their own defects and program themselves to produce superintelligent artificial agents that far surpass human capacities in virtually every cognitive domain.
- Kurzweil (1999, 2011), and other futuroligists, postulate the inevitability of superintelligent artificial agents as the necessary result of the inexorable rate of progress in computational technology.
- They cite Moore's law for the exponential growth in the processing power of computer chips as the analogical basis for this claim.

Artificial Super Intelligence as an Existential Threat

- A number of scientists (Häggström (2016)) and philosophers (Bostrom (2014)) adopt the idea of the singularity as the basis for claiming that strong AI may pose an existential threat to humanity.
- They envisage the possibility that super intelligent artificial agents could function autonomously in pursuit of goals that lead them to subordinate or destroy the human species.
- Advocates of this view maintain that the likelihood of the singularity, and of the threats that it will generate is significant to the point that this scenario constitutes a real and present danger that computer scientists, technologists, and public policy makers need to address now.

Artificial Super Intelligence as an Existential Threat

- A number of scientists (Häggström (2016)) and philosophers (Bostrom (2014)) adopt the idea of the singularity as the basis for claiming that strong AI may pose an existential threat to humanity.
- They envisage the possibility that super intelligent artificial agents could function autonomously in pursuit of goals that lead them to subordinate or destroy the human species.
- Advocates of this view maintain that the likelihood of the singularity, and of the threats that it will generate is significant to the point that this scenario constitutes a real and present danger that computer scientists, technologists, and public policy makers need to address now.

Artificial Super Intelligence as an Existential Threat

- A number of scientists (Häggström (2016)) and philosophers (Bostrom (2014)) adopt the idea of the singularity as the basis for claiming that strong AI may pose an existential threat to humanity.
- They envisage the possibility that super intelligent artificial agents could function autonomously in pursuit of goals that lead them to subordinate or destroy the human species.
- Advocates of this view maintain that the likelihood of the singularity, and of the threats that it will generate is significant to the point that this scenario constitutes a real and present danger that computer scientists, technologists, and public policy makers need to address now.

The Golem Goes into the Paperclip Business

- To illustrate his concerns Bostrom (2014) suggests a case in which a super intelligent robotic agent is charged with the objective of optimising and maximising the production of paperclips.
- Acting without additional constraints than those imposed by its task the robot progressively turns humans, the biospehere, and the entire solar system into material for efficiently producing a massive collection of paperclips.
- Its actions are motivated not by hostility to people, but by the fact that it assigns them no value beyond an instrumental role in the performance of its task.

The Golem Goes into the Paperclip Business

- To illustrate his concerns Bostrom (2014) suggests a case in which a super intelligent robotic agent is charged with the objective of optimising and maximising the production of paperclips.
- Acting without additional constraints than those imposed by its task the robot progressively turns humans, the biospehere, and the entire solar system into material for efficiently producing a massive collection of paperclips.
- Its actions are motivated not by hostility to people, but by the fact that it assigns them no value beyond an instrumental role in the performance of its task.

The Golem Goes into the Paperclip Business

- To illustrate his concerns Bostrom (2014) suggests a case in which a super intelligent robotic agent is charged with the objective of optimising and maximising the production of paperclips.
- Acting without additional constraints than those imposed by its task the robot progressively turns humans, the biospehere, and the entire solar system into material for efficiently producing a massive collection of paperclips.
- Its actions are motivated not by hostility to people, but by the fact that it assigns them no value beyond an instrumental role in the performance of its task.

AI and Moore's Law

- Progress in software development generally, and in AI in particular, does not follow the exponential tractory of the increase in processing power for hardware.
- When Allen and Greaves (2011) raised this objection to the singularity thesis Kurzweil (2011) replied by invoking the case of thermodynamic laws, which accurately predict the behaviour of collections of gas particles, despite the fact that the behaviour of each particle appears random.
- The analogy is misplaced, as the conduct of groups of scientists is not more accessible to accurate prediction than that of the individual researchers in these groups.

AI and Moore's Law

- Progress in software development generally, and in AI in particular, does not follow the exponential tractory of the increase in processing power for hardware.
- When Allen and Greaves (2011) raised this objection to the singularity thesis Kurzweil (2011) replied by invoking the case of thermodynamic laws, which accurately predict the behaviour of collections of gas particles, despite the fact that the behaviour of each particle appears random.
- The analogy is misplaced, as the conduct of groups of scientists is not more accessible to accurate prediction than that of the individual researchers in these groups.

AI and Moore's Law

- Progress in software development generally, and in AI in particular, does not follow the exponential tractory of the increase in processing power for hardware.
- When Allen and Greaves (2011) raised this objection to the singularity thesis Kurzweil (2011) replied by invoking the case of thermodynamic laws, which accurately predict the behaviour of collections of gas particles, despite the fact that the behaviour of each particle appears random.
- The analogy is misplaced, as the conduct of groups of scientists is not more accessible to accurate prediction than that of the individual researchers in these groups.

Autonomy and Volition

- In order to pose the sort of threat that Häggström and Bostrom are concerned about, super intelligent agents require a high degree of volitional independence.
- Even assuming self-improvement over a range of cognitive functions, and the capacity to implement secondary goals, it is not clear how volitional independence will emerge.
- Bostrom's paperclip golem is a fictional version of such an agent, but it is not obvious how AI systems, current or future, could generate one.
- But unless one can give a well motivated account of how an autonomous super intelligent agent could be created by technological means that are plausibly achievable, the fictional agent remains a work of imagination (like the original golem) rather than a genuine prospect.

Autonomy and Volition

- In order to pose the sort of threat that Häggström and Bostrom are concerned about, super intelligent agents require a high degree of volitional independence.
- Even assuming self-improvement over a range of cognitive functions, and the capacity to implement secondary goals, it is not clear how volitional independence will emerge.
- Bostrom's paperclip golem is a fictional version of such an agent, but it is not obvious how AI systems, current or future, could generate one.
- But unless one can give a well motivated account of how an autonomous super intelligent agent could be created by technological means that are plausibly achievable, the fictional agent remains a work of imagination (like the original golem) rather than a genuine prospect.

Autonomy and Volition

- In order to pose the sort of threat that Häggström and Bostrom are concerned about, super intelligent agents require a high degree of volitional independence.
- Even assuming self-improvement over a range of cognitive functions, and the capacity to implement secondary goals, it is not clear how volitional independence will emerge.
- Bostrom's paperclip golem is a fictional version of such an agent, but it is not obvious how AI systems, current or future, could generate one.
- But unless one can give a well motivated account of how an autonomous super intelligent agent could be created by technological means that are plausibly achievable, the fictional agent remains a work of imagination (like the original golem) rather than a genuine prospect.

Autonomy and Volition

- In order to pose the sort of threat that Häggström and Bostrom are concerned about, super intelligent agents require a high degree of volitional independence.
- Even assuming self-improvement over a range of cognitive functions, and the capacity to implement secondary goals, it is not clear how volitional independence will emerge.
- Bostrom's paperclip golem is a fictional version of such an agent, but it is not obvious how AI systems, current or future, could generate one.
- But unless one can give a well motivated account of how an autonomous super intelligent agent could be created by technological means that are plausibly achievable, the fictional agent remains a work of imagination (like the original golem) rather than a genuine prospect.

Deep Learning

- The application of deep learning with multi hidden layer neural networks has produced breakthroughs in a number of important areas in recent years.
- These include speech recognition, image identification, image description, diagnostics, and automatic driving.
- Unlike other machine learning methods deep learning involves self-correction through iterated training and development passes (using, for example, back propagation) .
- It also uses a set of domain general learning algorithms whose parameters are tuned to particular domains through training.

Deep Learning

- The application of deep learning with multi hidden layer neural networks has produced breakthroughs in a number of important areas in recent years.
- These include speech recognition, image identification, image description, diagnostics, and automatic driving.
- Unlike other machine learning methods deep learning involves self-correction through iterated training and development passes (using, for example, back propagation) .
- It also uses a set of domain general learning algorithms whose parameters are tuned to particular domains through training.

Deep Learning

- The application of deep learning with multi hidden layer neural networks has produced breakthroughs in a number of important areas in recent years.
- These include speech recognition, image identification, image description, diagnostics, and automatic driving.
- Unlike other machine learning methods deep learning involves self-correction through iterated training and development passes (using, for example, back propagation) .
- It also uses a set of domain general learning algorithms whose parameters are tuned to particular domains through training.

Deep Learning

- The application of deep learning with multi hidden layer neural networks has produced breakthroughs in a number of important areas in recent years.
- These include speech recognition, image identification, image description, diagnostics, and automatic driving.
- Unlike other machine learning methods deep learning involves self-correction through iterated training and development passes (using, for example, back propagation) .
- It also uses a set of domain general learning algorithms whose parameters are tuned to particular domains through training.

Real AI Does Not Work That Way

- But even the most sophisticated deep learning driven system that we can extrapolate from current technology would not have the volitional autonomy or domain general intelligence that advocates of the singularity attribute to a super intelligent agent.
- All current AI technology is entirely task driven, and lacking volition.
- It is not in any way obvious how even radical improvements or modifications of current systems could yield the sorts of agents that the singularity thesis assumes.
- So the realisation of such an agent would require an entirely different sort of AI technology than any that we have created, or could envisage now.

Real AI Does Not Work That Way

- But even the most sophisticated deep learning driven system that we can extrapolate from current technology would not have the volitional autonomy or domain general intelligence that advocates of the singularity attribute to a super intelligent agent.
- All current AI technology is entirely task driven, and lacking volition.
- It is not in any way obvious how even radical improvements or modifications of current systems could yield the sorts of agents that the singularity thesis assumes.
- So the realisation of such an agent would require an entirely different sort of AI technology than any that we have created, or could envisage now.

Real AI Does Not Work That Way

- But even the most sophisticated deep learning driven system that we can extrapolate from current technology would not have the volitional autonomy or domain general intelligence that advocates of the singularity attribute to a super intelligent agent.
- All current AI technology is entirely task driven, and lacking volition.
- It is not in any way obvious how even radical improvements or modifications of current systems could yield the sorts of agents that the singularity thesis assumes.
- So the realisation of such an agent would require an entirely different sort of AI technology than any that we have created, or could envisage now.

Real AI Does Not Work That Way

- But even the most sophisticated deep learning driven system that we can extrapolate from current technology would not have the volitional autonomy or domain general intelligence that advocates of the singularity attribute to a super intelligent agent.
- All current AI technology is entirely task driven, and lacking volition.
- It is not in any way obvious how even radical improvements or modifications of current systems could yield the sorts of agents that the singularity thesis assumes.
- So the realisation of such an agent would require an entirely different sort of AI technology than any that we have created, or could envisage now.

Mapping the Brain

- Kurzweill (1999, 2011) suggests that one way in which we might achieve the singularity is by mapping all of the neural connections of the human brain, and implementing an artificial version of this neural mechanism.
- But a full mapping of the brain will not, in itself, produce an understanding of how it works computationally.
- Even if it did, it is not obvious that constructing such a device will produce the sort of super intelligent artificial agent that he envisages.
- We have decoded the genomes of basic organisms, but we are far from reproducing their biological properties and patterns of behaviour through synthesized genes.

Mapping the Brain

- Kurzweill (1999, 2011) suggests that one way in which we might achieve the singularity is by mapping all of the neural connections of the human brain, and implementing an artificial version of this neural mechanism.
- But a full mapping of the brain will not, in itself, produce an understanding of how it works computationally.
- Even if it did, it is not obvious that constructing such a device will produce the sort of super intelligent artificial agent that he envisages.
- We have decoded the genomes of basic organisms, but we are far from reproducing their biological properties and patterns of behaviour through synthesized genes.

Mapping the Brain

- Kurzweill (1999, 2011) suggests that one way in which we might achieve the singularity is by mapping all of the neural connections of the human brain, and implementing an artificial version of this neural mechanism.
- But a full mapping of the brain will not, in itself, produce an understanding of how it works computationally.
- Even if it did, it is not obvious that constructing such a device will produce the sort of super intelligent artificial agent that he envisages.
- We have decoded the genomes of basic organisms, but we are far from reproducing their biological properties and patterns of behaviour through synthesized genes.

Mapping the Brain

- Kurzweill (1999, 2011) suggests that one way in which we might achieve the singularity is by mapping all of the neural connections of the human brain, and implementing an artificial version of this neural mechanism.
- But a full mapping of the brain will not, in itself, produce an understanding of how it works computationally.
- Even if it did, it is not obvious that constructing such a device will produce the sort of super intelligent artificial agent that he envisages.
- We have decoded the genomes of basic organisms, but we are far from reproducing their biological properties and patterns of behaviour through synthesized genes.

From Possibility to Risk Assessment

- It is certainly the case that the singularity, and malevolent super agents are logically possible events.
- Until detailed arguments are presented showing how such events would come about, they are not more than that.
- They join the ranks of other remotely possible, but unlikely occurrences, like the sudden appearance of black holes in the vicinity of the Earth, and extra-terrestrial invasion.
- When assessing the risks associated with possible events, we prioritise them according to the likelihood of these events taking place in the foreseeable future.

From Possibility to Risk Assessment

- It is certainly the case that the singularity, and malevolent super agents are logically possible events.
- Until detailed arguments are presented showing how such events would come about, they are not more than that.
- They join the ranks of other remotely possible, but unlikely occurrences, like the sudden appearance of black holes in the vicinity of the Earth, and extra-terrestrial invasion.
- When assessing the risks associated with possible events, we prioritise them according to the likelihood of these events taking place in the foreseeable future.

From Possibility to Risk Assessment

- It is certainly the case that the singularity, and malevolent super agents are logically possible events.
- Until detailed arguments are presented showing how such events would come about, they are not more than that.
- They join the ranks of other remotely possible, but unlikely occurrences, like the sudden appearance of black holes in the vicinity of the Earth, and extra-terrestrial invasion.
- When assessing the risks associated with possible events, we prioritise them according to the likelihood of these events taking place in the foreseeable future.

From Possibility to Risk Assessment

- It is certainly the case that the singularity, and malevolent super agents are logically possible events.
- Until detailed arguments are presented showing how such events would come about, they are not more than that.
- They join the ranks of other remotely possible, but unlikely occurrences, like the sudden appearance of black holes in the vicinity of the Earth, and extra-terrestrial invasion.
- When assessing the risks associated with possible events, we prioritise them according to the likelihood of these events taking place in the foreseeable future.

The Problem of Faulty Design

- It is necessary to distinguish the prospect of malevolent super agents from the likelihood that an AI technology will cause unintended harm through faulty design.
- The latter is a very real concern, which requires constant attention.
- It is not specific to AI, or to computer technology.
- It attaches to every artefact, from simple tools and weapons to advanced communications systems.
- The more complex a system is, the more challenging is the task of filtering out its design faults.

The Problem of Faulty Design

- It is necessary to distinguish the prospect of malevolent super agents from the likelihood that an AI technology will cause unintended harm through faulty design.
- The latter is a very real concern, which requires constant attention.
- It is not specific to AI, or to computer technology.
- It attaches to every artefact, from simple tools and weapons to advanced communications systems.
- The more complex a system is, the more challenging is the task of filtering out its design faults.

The Problem of Faulty Design

- It is necessary to distinguish the prospect of malevolent super agents from the likelihood that an AI technology will cause unintended harm through faulty design.
- The latter is a very real concern, which requires constant attention.
- It is not specific to AI, or to computer technology.
- It attaches to every artefact, from simple tools and weapons to advanced communications systems.
- The more complex a system is, the more challenging is the task of filtering out its design faults.

The Problem of Faulty Design

- It is necessary to distinguish the prospect of malevolent super agents from the likelihood that an AI technology will cause unintended harm through faulty design.
- The latter is a very real concern, which requires constant attention.
- It is not specific to AI, or to computer technology.
- It attaches to every artefact, from simple tools and weapons to advanced communications systems.
- The more complex a system is, the more challenging is the task of filtering out its design faults.

The Problem of Faulty Design

- It is necessary to distinguish the prospect of malevolent super agents from the likelihood that an AI technology will cause unintended harm through faulty design.
- The latter is a very real concern, which requires constant attention.
- It is not specific to AI, or to computer technology.
- It attaches to every artefact, from simple tools and weapons to advanced communications systems.
- The more complex a system is, the more challenging is the task of filtering out its design faults.

AI as the Driver for Large Scale Automation

Permanent Large Scale Unemployment

- In previous industrial revolutions the loss of employment in one domain was compensated for by the opening up of production and services in new areas.
- The future wave of automation may well be different in kind.
- It could spread through all areas of the economy, rendering human labour unnecessary across a wide range of domains.
- This may result in large scale unemployment as a permanent feature of a technologically advanced economy.

Permanent Large Scale Unemployment

- In previous industrial revolutions the loss of employment in one domain was compensated for by the opening up of production and services in new areas.
- The future wave of automation may well be different in kind.
- It could spread through all areas of the economy, rendering human labour unnecessary across a wide range of domains.
- This may result in large scale unemployment as a permanent feature of a technologically advanced economy.

Permanent Large Scale Unemployment

- In previous industrial revolutions the loss of employment in one domain was compensated for by the opening up of production and services in new areas.
- The future wave of automation may well be different in kind.
- It could spread through all areas of the economy, rendering human labour unnecessary across a wide range of domains.
- This may result in large scale unemployment as a permanent feature of a technologically advanced economy.

Permanent Large Scale Unemployment

- In previous industrial revolutions the loss of employment in one domain was compensated for by the opening up of production and services in new areas.
- The future wave of automation may well be different in kind.
- It could spread through all areas of the economy, rendering human labour unnecessary across a wide range of domains.
- This may result in large scale unemployment as a permanent feature of a technologically advanced economy.

Radical Disparity in the Distribution of Wealth

- If the small group of people who design, produce, and market AI driven robotic systems control production and services, then there will be a very sharp division of wealth between this technology sector and the rest of the population.
- Labour will become an increasingly unreliable source of income, as the demand for it declines.
- This situation will pose a serious threat to social cohesion.

Radical Disparity in the Distribution of Wealth

- If the small group of people who design, produce, and market AI driven robotic systems control production and services, then there will be a very sharp division of wealth between this technology sector and the rest of the population.
- Labour will become an increasingly unreliable source of income, as the demand for it declines.
- This situation will pose a serious threat to social cohesion.

Radical Disparity in the Distribution of Wealth

- If the small group of people who design, produce, and market AI driven robotic systems control production and services, then there will be a very sharp division of wealth between this technology sector and the rest of the population.
- Labour will become an increasingly unreliable source of income, as the demand for it declines.
- This situation will pose a serious threat to social cohesion.

The Need for Public Planning and Intervention

- It is unclear how even highly regulated free market economies, in their current form, can cope with the economic crises that pervasive automation may generate.
- Large scale unemployment will emerge as an instance of market failure, created by the very technology that drives the market itself.
- It is difficult to see how to solve this problem without public intervention to achieve wide spread redistribution of wealth.
- If this does turn out to be the case, an important question is whether redistribution should provide a guaranteed income, or subsidise employment in services for which the market does not provide adequate demand.

The Need for Public Planning and Intervention

- It is unclear how even highly regulated free market economies, in their current form, can cope with the economic crises that pervasive automation may generate.
- Large scale unemployment will emerge as an instance of market failure, created by the very technology that drives the market itself.
- It is difficult to see how to solve this problem without public intervention to achieve wide spread redistribution of wealth.
- If this does turn out to be the case, an important question is whether redistribution should provide a guaranteed income, or subsidise employment in services for which the market does not provide adequate demand.

The Need for Public Planning and Intervention

- It is unclear how even highly regulated free market economies, in their current form, can cope with the economic crises that pervasive automation may generate.
- Large scale unemployment will emerge as an instance of market failure, created by the very technology that drives the market itself.
- It is difficult to see how to solve this problem without public intervention to achieve wide spread redistribution of wealth.
- If this does turn out to be the case, an important question is whether redistribution should provide a guaranteed income, or subsidise employment in services for which the market does not provide adequate demand.

The Need for Public Planning and Intervention

- It is unclear how even highly regulated free market economies, in their current form, can cope with the economic crises that pervasive automation may generate.
- Large scale unemployment will emerge as an instance of market failure, created by the very technology that drives the market itself.
- It is difficult to see how to solve this problem without public intervention to achieve wide spread redistribution of wealth.
- If this does turn out to be the case, an important question is whether redistribution should provide a guaranteed income, or subsidise employment in services for which the market does not provide adequate demand.

Conclusions

- The emergence of super intelligent robotic agents that pose a threat to humanity may be a logical possibility, but it does not seem likely at any time in the foreseeable future.
- Current AI technology, and any technology derivable from it, do not support this scenario in any plausible way.
- Widespread automation in large sections of the economy does appear likely, and it carries with it serious economic and social problems.
- Public planning and intervention are urgently needed to deal with these issues.

Conclusions

- The emergence of super intelligent robotic agents that pose a threat to humanity may be a logical possibility, but it does not seem likely at any time in the foreseeable future.
- Current AI technology, and any technology derivable from it, do not support this scenario in any plausible way.
- Widespread automation in large sections of the economy does appear likely, and it carries with it serious economic and social problems.
- Public planning and intervention are urgently needed to deal with these issues.

Conclusions

- The emergence of super intelligent robotic agents that pose a threat to humanity may be a logical possibility, but it does not seem likely at any time in the foreseeable future.
- Current AI technology, and any technology derivable from it, do not support this scenario in any plausible way.
- Widespread automation in large sections of the economy does appear likely, and it carries with it serious economic and social problems.
- Public planning and intervention are urgently needed to deal with these issues.

Conclusions

- The emergence of super intelligent robotic agents that pose a threat to humanity may be a logical possibility, but it does not seem likely at any time in the foreseeable future.
- Current AI technology, and any technology derivable from it, do not support this scenario in any plausible way.
- Widespread automation in large sections of the economy does appear likely, and it carries with it serious economic and social problems.
- Public planning and intervention are urgently needed to deal with these issues.