UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE
DÉPARTEMENT DES LANGUES

Examen écrit — Session 1 — Semestre 1 — 2019-2020

Langue : **Anglais**	Niveau : **4**	Date : 11 janvier 2020	Durée : 1h30

The Rise of Robots: forget evil AI – the real risk is far more insidious

Olivia Solon, Tuesday 30th August 2016, *The Guardian*

When we look at the rise of artificial intelligence, it's easy to get carried away with dystopian visions of sentient machines that rebel against their human creators. Fictional baddies such as *Terminator*'s Skynet or Hal from *2001: A Space Odyssey* have a lot to answer for.

However, the real risk posed by AI – at least in the near term – is much more insidious. It's far more likely that robots would inadvertently harm or frustrate humans while carrying out our orders than they would become conscious and rise up against us. In recognition of this, the University of California, Berkeley has this week launched a center to focus on building people-pleasing AIs.

The Center for Human-Compatible Artificial Intelligence, launched this week with $5.5m in funding from the Open Philanthropy Project, is led by computer science professor and artificial intelligence pioneer Stuart Russell. He's quick to dispel any "unreasonable and melodramatic" comparisons to the threats posed in science fiction.

"The risk doesn't come from machines suddenly developing spontaneous malevolent consciousness," he said. "It's important that we're not trying to prevent that from happening because there's absolutely no understanding of consciousness whatsoever."

Russell is well known in the artificial intelligence community and in 2015 penned an open letter calling for researchers to look beyond the goal of simply making AI more capable and powerful to think about maximizing its social benefit. The letter has been signed by more than 8,000 scientists and entrepreneurs including physicist Stephen Hawking, entrepreneur Elon Musk and Apple co-founder Steve Wozniak.

"The potential benefits [of AI research] are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty is not improbable," the letter reads.

"Because of the great potential of AI, it is important to research how to get the benefits while avoiding potential pitfalls."

It's precisely this thinking that underpins the new center.

Up until now, AI has primarily been applied to very limited contexts such as playing Chess or Go or recognizing objects in images, where there isn't much scope for the system to do much damage. As they start to make decisions on our behalf within the real world, the stakes are much higher.

"As soon as you put things in the real world, with self-driving cars, digital assistants … as soon as they buy things on your behalf, turn down appointments, then they have to align with human values," Russell said.

He uses autonomous vehicles to illustrate the type of problem the center will try to solve. Someone building a self-driving car might instruct it never to go through a red light, but the machine might then hack into the traffic light control system so that all of the lights are changed to green. In this case the car would be obeying orders but in a way that humans didn't expect or intend. […]

"Even when you think you've put limits to what an AI system can do it will tend to find loopholes just as we do with our tax laws. You want an AI system that isn't motivated to find loopholes," Russell said.

"The problem isn't consciousness, but competence. You make machines that are incredibly competent at achieving objectives and they will cause accidents in trying to achieve those objectives."

To address this, Russell and his colleagues at the center propose making AI systems that observe human behavior and try to work out what the human's objective is, then behave accordingly and learn from mistakes. So instead of trying to give the machine a long list of rules to follow, the machine is told that its main objective is to do what the human wants them to do.

It sounds simple, but it's not how engineers have been building systems for the past 50 years.

But if AI systems can be designed to learn from humans in this way, it should ensure that they remain under human control even when they develop capabilities that exceed our own.

In addition to watching humans directly using cameras and other sensors, robots can learn about us by reading history books, legal documents, novels, newspaper stories as well as by watching videos and movies. From this they can start to build up an understanding of human values.

It won't be easy for machines. "People are irrational, inconsistent, weak-willed, computationally limited, heterogenous and sometimes downright evil," Russell said.

"Some are vegetarians and some really like a nice juicy steak. And the fact that we don't behave anything close to perfectly is a serious difficulty."

A. Reading (9 marks)

1) Find the **synonyms** of the following words in the text. The words below appear **in the order of the text**. (/3)

➔ **Between lines 1 and 27**

a. conscious, sensible

b. to dismiss, to reject

c. dangers, problems

➔ **From lines 29 to 42**

d. for us, instead of us

e. risks

f. ambiguities/omissions in legislation

2) **Based on Part 1 (ll.1-28)**, explain why we shouldn't be afraid of a robot rebellion. Why should we welcome AI instead? Answer **in your words**, using terms marking **opposition and contrast.** (/3)

3) **Use the information from Part 2 (ll.29-end) to sum up** what "people-pleasing AIs" (l.9) are. How would you define them? What is the purpose of building them? How would they be built? (/3)

B. Grammar (3 marks)

These sentences adapted from the text are in the **passive voice**. Change them into the **active voice**, paying attention to word order and tenses.

1) The Center for Human-Compatible AI is led by computer science professor Stuart Russell.

2) The letter has been signed by more than 8,000 scientists.

3) Human intelligence will be magnified by the tools of AI.

C. Writing (8 marks)

Answer **ONE** of the questions below in 250 words (+/- 10%). Make sure you write a well-organised essay and use your own words.

1) Apart from the “eradication of disease and poverty” (l.24), name **ONE** huge potential benefit of AI. Justify your choice.

OR

2) “Self-driving cars, digital assistants” (l.33)… Will AI end up eroding humans’ ability to think for themselves and take independent action?

OR

3) If the threats dealt with in science fiction are “unreasonable and melodramatic” (l.12), why is the concept of evil AI so popular in films and series?