WHAT DISTINGUISHES A GOOD AI FROM A BAD AI?

SAARLAND UNIVERSITY

Name: Priyanka Upadhyay (Master's DSAI student)
Matriculation: 2581714
Email ID: s8prupad@stud.uni-saarland.de

MOTIVATION

- You will never know what will be the consequences of the misfortune; or, you never know the consequences of good fortune.
- Can we actually do something right or wrong?
- As AI infiltrates more of our experiences and organizations, it's important to recognize not only its many benefits but its unintended consequences as well.
- AI protects us from known and unknown threats, helps us connect to one another, and provides better answers faster and cheaper than humans do. And, of course, it's great that AI frees us from routine tasks such as reading a map. But are we recognizing and addressing the loss of human expertise that accompanies that new freedom? - **Bain & Company, Inc.** [17]

Can we code right and wrong?

CONTENTS

1. Good And Evil: Views Of The Philosophers [1]

“What is good and what is evil?”, Many Philosophers of all ages have given their thought over this question.

Questions:

A few basic questions need to be answered in order to arrive at some satisfactory answer. They are:

I. Are good and evil absolute or are they relative to the conditions associated with time and place?

II. Is the concept of good and evil imbued in the nature of man or has he been given divine guidance?

III. If good and evil are independent, do they have the same creator?

IV. If the knowledge of good and evil is instinctive, there should be uniformity of thought between various nations, religions and groups; but there are vast differences among them in almost every aspect. What are the reasons?

Answers:

Those above questions have been thought over by **philosophers and thinkers** of all ages. Here are few names of them:-

- Heraclitus(535-475 BC)
- Soctrates(470-390 BC)
- Democritus(460-370 BC)
- Sophists Philosophy
- Plato(428-348 BC)
- Aristotle(384-322 BC)
- Saint Augustine(354-430 BC)
- Epicurean and Stoic Philosophy
- Philo(50-30 BC)
- Peter Abelard (1079-1142 AD)
- Thomas Aquinas (1227-1274 AD)
- Meister Eckhart (1260-1327 AD)
- Thomas Hobbes (1588-1679 AD)
- Descartes (1596-1650 AD)
- Spinoza (1632-1677 AD)
- John Locke (1632-1704 AD)
- Liebnitz (1646-1716 AD)
- Richard Cumberland (1631-1718 AD)
- Immannual Kant (1724-1804 AD)
- Arthur Shopenhauer (1788-1860 AD)
- John Stuart Mill (1806-1873 AD)
- Jeremy Bentham (1874-1832 AD)
- Herbert Spencer (1820-1903 AD)
- William James (1842-1910 AD) & John Dewey (1859-1952 AD)

We are going to briefly discuss their views. However, I am going to mention only those philosophers's views who has left a deep impact upon philosophical thought.

Views of Philosophers:

1. **Democritus (460-370 BC):**

He said: "A good man is not one who does good, but who always wants to do good."

2. **Sophists Philosophy:**

- One of the sophist, Protagoras (who was generally regarded as the first of these professional sophists), considered **"Man is the measure of all things,**". Everybody has the right to determine for himself what is good and what is evil for them.
- Thrasymachus and Callicles said that "**There are no moral laws, no all-inclusive principles to do right and wrong. He is free to live as he desires and to get what he wants by any means possible."** However, since the outcome was moral anarchy, pure individualism and selfishness**, Callicles went as far as saying: "To hell with morality, this has been propounded by the weak to debilitate the power of the strong."**

3. Soctrates(470-390 BC):

Socrate's emphasis on self-realization and innate knowledge. He said, “No man is voluntarily bad. He turns bad when he does not know what is good and what is evil. If he knew what is good, he was sure to choose it.”

4. Plato(428-348 BC):

He thought that man is endowed with the knowledge of good and evil before coming to this world. This knowledge existed in his soul but during the period between his creation and his descent in this world, he forgot most of the things. These forgotten things can be recollected either by wise sermons or through meditation on nature. Experience also helps in recollection of the forgotten. All good and evil is innate in man.

5. **Aristotle(384-322 BC):**

He thought that reason is the greatest bounty of God, and called it the `Divine Spark'. If man uses his reason and other capabilities properly, he can attain self-realization after which he hardly needs any measure for good and evil.

6. **Philo(50-30 BC):**

Philo thought that the spiritual part of man, his mind or soul, is the seat of good, and his body, the material part, is the seat of evil. Consequently, when the soul is incorporated in the body it suffers a fall from divine perfection and becomes predisposed to evil.

7. Saint Augustine(354-430 BC):

The early Christian thinkers thought that God had given man a good nature, but he had turned away from God to the flesh i.e, the body**.**

Saint Augustine, the greatest of the Christian thinkers, thought that **God is all good, all perfection. He cannot be the creator of evil. `How then to account for evil in a world created by an all-good God?'** To solve this problem, Saint Augustine said that everything in the universe is good; even that which appears to be evil is actually good in as much as it fitted into the whole pattern of the universe.

Examples: Flowers of different colours are necessary to the beauty of a garden and every flower is good in its own place adding to the beauty of the garden.

Similarly, the evil which is found in the world is there to make the whole good. It looks evil only when one sees the dark spots broken away from the whole picture but when seen in the picture they add to its beauty. **If we fit evil in the whole system of the universe, it would look good and beautiful.**

Critical examination of philosophical views:

The foregoing discussion summarizes the views of philosophers and thinkers who thought over the problem of good and evil and tried to answer the questions mentioned earlier. A critical examination of these views follows:

1. Philosophy of relativeness:

Philiosophers who uphold this philosophy are those who have exemplified good and evil with musical notes or dark and red shades of a picture. Heraclitus and St Augustine are in this category. For them, the high and low notes in a symphony or shades in a picture increase its attractiveness. Similarly good and evil are essential for the world's beauty and charm.

2. Criterion of Intention:

Similarly, the views of the philosophers who maintain that an act is neither good nor bad in itself but intention makes it so are equally incorrect. Mere intention cannot make a bad act good. At the most a bad act performed in good faith can be excused but it cannot be classified as a good act. Therefore, intention cannot be made the basis of determining good and evil. This view is without a rationale.

3. Criterion of Pleasure and Happiness:

The philosophers who consider pleasure and happiness to be the criterion of good are also far from the truth. The criterion of pleasure and happiness is baseless. No single measure can be laid down for pleasure and happiness.

4. The theory of Divine Intuition:

Great philosophers like Socrates, Plato, Aristotle, Kant, John Locke and Liebnitz, all of them, considered good and evil to be independent. The foundation, of both already exists in man's nature. Conception of moral laws is innate in man; There is a Divine Spark within him to guide him. Man often forgets the moral laws and needs to be remained.

Existence of Evil:

Consider the question, `if good and evil were independent, are their creators also independent?'. Most of the philosophers have not discussed this question.

Those who thought good and evil to be relative, dumped the question itself.

Polytheists consider gods of good and evil to be independent of each other. Christian thinkers known as Apologists also answered the question as the polytheists did. According to their representatives, Thomas Aquinas and St. Augustine, God is pure good and He cannot be the Creator of evil.

2. Good and Evil: according to religion

- **In cultures with Manichaean and Abrahamic religious influence:**

Evil is usually perceived as the dualistic antagonistic opposite of good, in which good should prevail and evil should be defeated[3].

- **In cultures with Buddhist spiritual influence:**

Both good and evil are perceived as part of an - antagonistic duality that itself must be overcome through achieving emptiness in the sense of recognition of good and evil. – Fig 2

Fig 1 - In many religions, angels are considered good beings. In the Judeo-Christian tradition, God —being the creator of all life —is seen as the personification of good.

Fig 2 - One of the five paintings of Extermination of Evil portrays Sendan Kendatsuba, one of the eight guardians of Buddhist law, banishing evil.

3. But in reality, Does anyone know what is Good or Bad?

- Good and Bad are relative terms. it's so completely different for every individual. properly speaking from a philosophical facet, everybody has a perspective of good or bad applied to their actions (Right or Wrong). This can be the conscience.
- Freud regarded conscience as originating psychologically from the growth of civilizations. He considered that it suppressed the expression of aggression and direct the aggressive energy to better, healthy proper channels.

  So everybody really does know consciously or sub-consciously what is good or bad.

- "No man's conscience can tell him the right of another man." - **Samuel Johnson**

4. A Farmer case[6]

I'll tell you a story about a Chinese farmer:

- There was once a Chinese farmer who lost a horse because it ran away. Later that evening, all the neighbors came over and said "That's too bad", the farmer said "Maybe".
- The next day, the horse came back and brought several wild horses with it. The neighbors said "Well that's great isn't it?!", the farmer said "Maybe".
- The next day, his son, in an attempt to tame one of these horses, fell and broke his leg. The neighbors came over and said "That's too bad", and the farmer said "Maybe".
- The next day, the conscription officers came around looking for people for the army. They looked at his son and rejected him because of his broken leg. The neighbors came by and said "Well that's great isn't it?!", the farmer replied "Maybe". — Alan Watts

The point being, it is really impossible to tell if anything within life is really good or bad. Because you will never know what will be the consequences of the misfortune; or, you never know the consequences of good fortune.

5. If good or bad exist, how does it work within the society? [7]

The golden rule is:

- "As you would have people do to you, do to them; and what you dislike to be done to you, don't do to them" - Prophet Mohamed
- "That which is unfavorable to us, do not do that to others" - Padmapuraana, shrushti 19/357–358 (Puranas)
- "Hurt not others in ways that you yourself would find hurtful" - Buddha - Udanavarga, 5:18
- "Do not do unto others whatever is injurious to yourself" – Shayast-na-Shayast 13.29 (Zoroastrianism)
- "Do to others what you want them to do to you" - Mathew

6. What is AI?

- Now, In technical world, we do have technologies which can solve our many problems. Artificial intelligence is one of them and all of the theories which we have discussed till now on Good and Bad were to build up the ethics and morality in AI – so that machine can take right decision for human kind.

- The term artificial intelligence was first coined by John McCarthy in 1956 when he held the first academic conference on the subject. But the journey to understand if machines can truly think began much before that. In Vannevar Bush's seminal work As We May Think [Bush45] **he proposed a system which amplifies people's own knowledge and understanding**. Five years later Alan Turing wrote a paper on the notion of machines being able to simulate human beings and the ability to do intelligent things [8].

7. Philosophy of Good and Bad Artificial Intelligence

Philosophy of Artificial Intelligence:

The philosophy of artificial intelligence attempts to answer such questions as follows:

- Can a machine act intelligently? Can it solve any problem that a person would solve by thinking?
- Are human intelligence and machine intelligence the same? Is the human brain essentially a computer?
- Can a machine have a mind, mental states, and consciousness in the same sense that a human being can? Can it feel how things are?

Good Artificial Intelligence (How can AI be good for society?):

- Artificial intelligence can dramatically improve the efficiencies of our workplaces and can augment the work humans can do. **When AI takes over repetitive tasks,** it frees up the human workforce to do work they are better equipped for—tasks that involve creativity and empathy among others.

Bad Artificial Intelligence (How can AI be Dangerous for society ?):

The AI is programmed to do something devastating:

- Autonomous weapons are artificial intelligence systems that are programmed to kill. In the hands of the wrong person, these weapons could easily cause mass casualties.
- To avoid being thwarted by the enemy, these weapons would be designed to be extremely difficult to simply “turn off,”.

The AI is programmed to do something beneficial, but it develops a destructive method for achieving its goal:

This can happen whenever we fail to fully align the AI’s goals with ours, which is strikingly difficult.

- If you ask an obedient intelligent car to take you to the airport as fast as possible, it might get you there chased by helicopters and covered in vomit, doing not what you wanted but literally what you asked for.

.

8. Impact on society : Benefits and Risks of AI[9]

"Everything we love about civilization is a product of intelligence, so amplifying our human intelligence with artificial intelligence has the potential of helping civilization flourish like never before – as long as we manage to keep the technology beneficial."

-**Max Tegmark**, President of the Future of Life Institute

Benefits of AI:

- With better monitoring and diagnostic capabilities, artificial intelligence can dramatically influence healthcare.
- The way we uncover criminal activity and solve crimes will be enhanced with artificial intelligence. Facial recognition technology is becoming just as common as fingerprints.

- **Positive impacts of AI**[10]:
  - Picture phone and teleconference
  - Television and video
  - Music
  - Shopping
  - Online banking and financial services
  - Reservations
  - Medical advice
  - Video games and Other games (e.g., gambling, chess etc.)
  - News, sports and weather reports
  - Access to data banks.

- **Wonderful Examples Of Using Artificial Intelligence (AI) For Good [11]:**
  - Cancer Screening
  - Tools for disable People
  - Climate change
  - Wildlife conservation
  - Combat World Hunger
  - Reduce Inequality and Poverty
  - Assess Medical Images
  - Spot “Fake News”

Risks of AI:

- **Negative impacts of AI[12]:**
  - AI Bias
  - Loss of Certain Jobs
  - A shift in Human Experience
  - Accelerated Hacking

- **Dangerous Examples Of Using Artificial Intelligence (AI) For Bad [13]:**
  - Uber self-driving car kills a pedestrian
  - IBM Watson comes up short in healthcare
  - Chinese billionaire's face identified as jaywalker
  - Amazon AI recruiting tool is gender biased
  - DeepFakes reveals AI's unseemly side
  - Google Photo confuses skier and mountain
  - AI World Cup 2018 predictions almost all wrong
  - Startup claims to predict IQ from faces

9. Challenges[14]

- Artificial intelligence will definitely cause our workforce to evolve. The alarmist headlines emphasize the loss of jobs to machines, but the real challenge is for humans to find their passion with new responsibilities that require their uniquely human abilities. According to PwC, 7 million existing jobs will be replaced by AI in the UK from 2017-2037, but 7.2 million jobs could be created. This uncertainty and the changes to how some will make a living could be challenging.

- Another issue is ensuring that **AI doesn't become so proficient at doing the job it was designed to do that it crosses over ethical or legal boundaries. While the original intent and goal of the AI is to benefit humanity, if it chooses to go about achieving the desired goal in a destructive (yet efficient way) it would negatively impact society.** The AI algorithms must be built to align with the overarching goals of humans.

10. How can we distinguish a Good AI from a Bad AI? and how it can be improved?

We had better be quite sure that the purpose put into the machine is the purpose which we really desire." – Norbert Wiener, 1960

When researcher were thinking to develop this idea and few were already developed then they had **"The value alignment problem"** and according to this – we put the wrong objective into the machine, though we wanted the same objective output but we did not give specific instruction to machine and it turns out to be in wrong result.

Example: In classic AI: if you put an objective into the machine, even something "fetch the coffee", the machine says to itself "Well, How might I fell to fetch the coffee?", Someone might switch me off, okay I have to take step to prevent that I will disable my "off" switch. I will do anything to defend myself again interference with this objective which I have been given. So this single minded pursuit in a very defensive mode of an objective that is, in fact, not aligned with the true objective of the human race that's the problem which we face.

Classic AI:

Isaac Asimov's "Three Laws of Robotics"[15]:

- **First Law:** A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- **Second Law:** A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
- **Third Law:** A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Human compatible AI:

Stuart Russell's "3 Principle for creating safer AI"[16] :

So to solve human race problem, we need to follow some principle to convert Classic AI into human compatible AI:-

- **First Law:** The robot's only objective is to maximize the realization of human objectives/values.

And in this way it actually violates the Asimov's third law, that the robot has to protect its own existence. In Stuart's law it has no interest in preserving its own existence whatsoever.

- **Second Law:** The robot is initially uncertain about what those human values are.

So human race problem is solved here as robot takes values from third law

- **Third Law:** Human behavior provides information about human values.

So our own choices reveal information about what it is we prefer our lives to be like

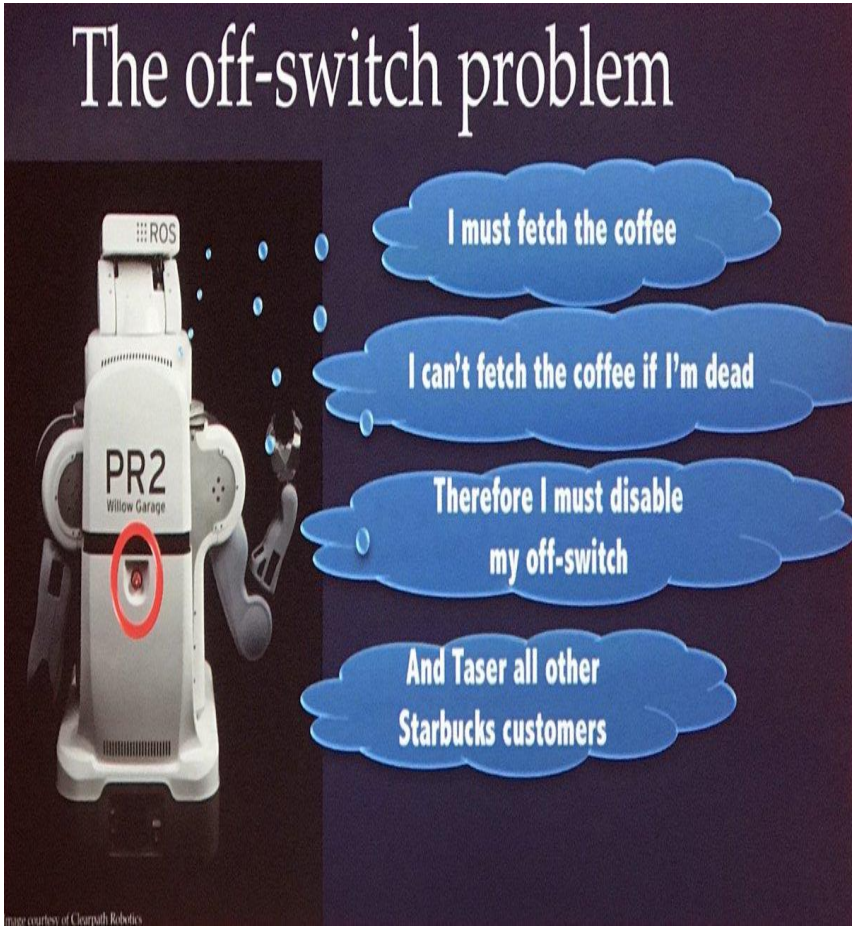

Fig 3. Classic AI [16]

Fig 4. Human Compatible AI [16]
(The machine will let Human switch it off according to first and second principle and when machine is "off", then third law will be followed)

So from Stuart ‘s third Law, Human behavior provides information about human values. Where human values differ from each other -

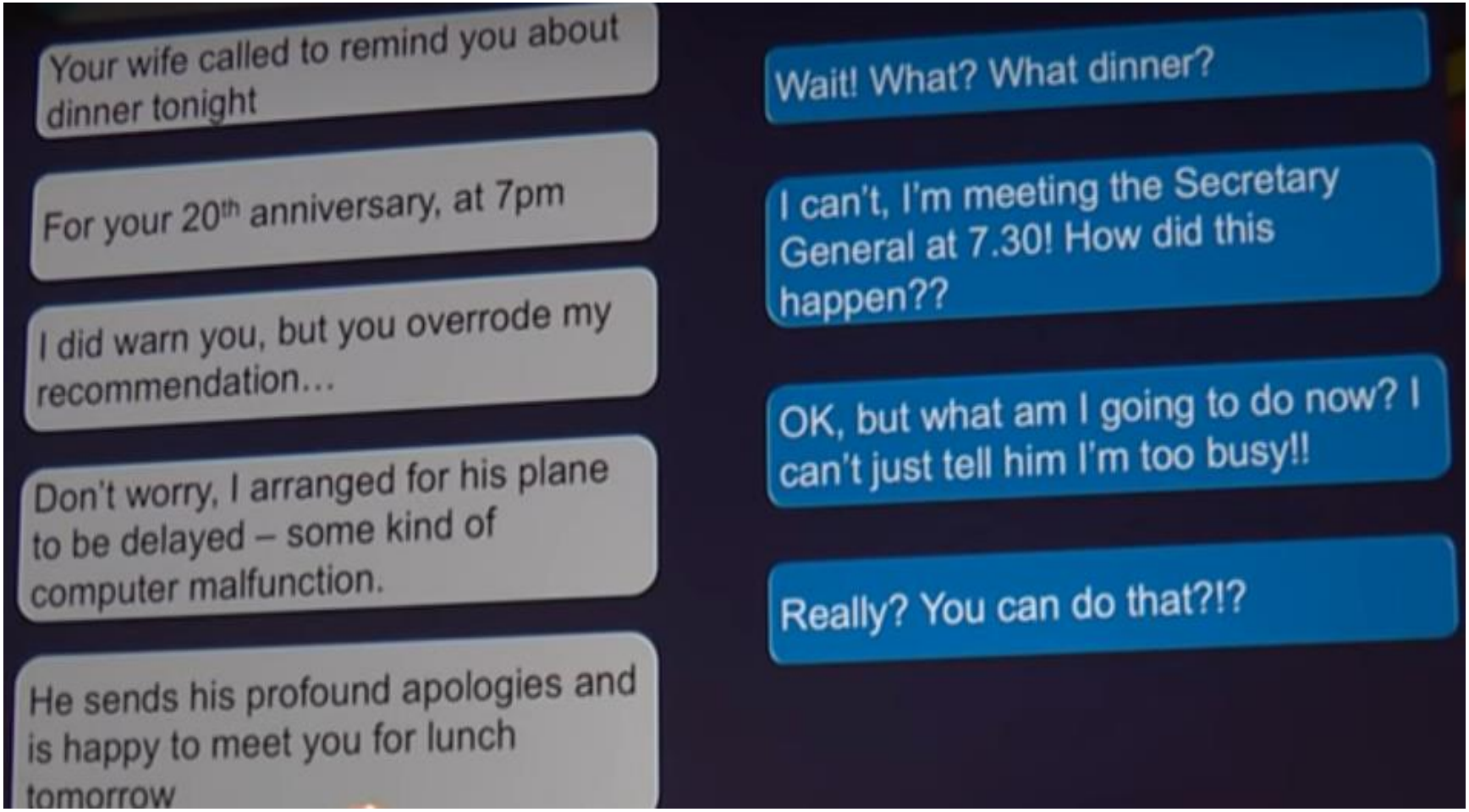

Fig 5. Giving priority to wife objective- Happy wife ! Happy Life! [16]

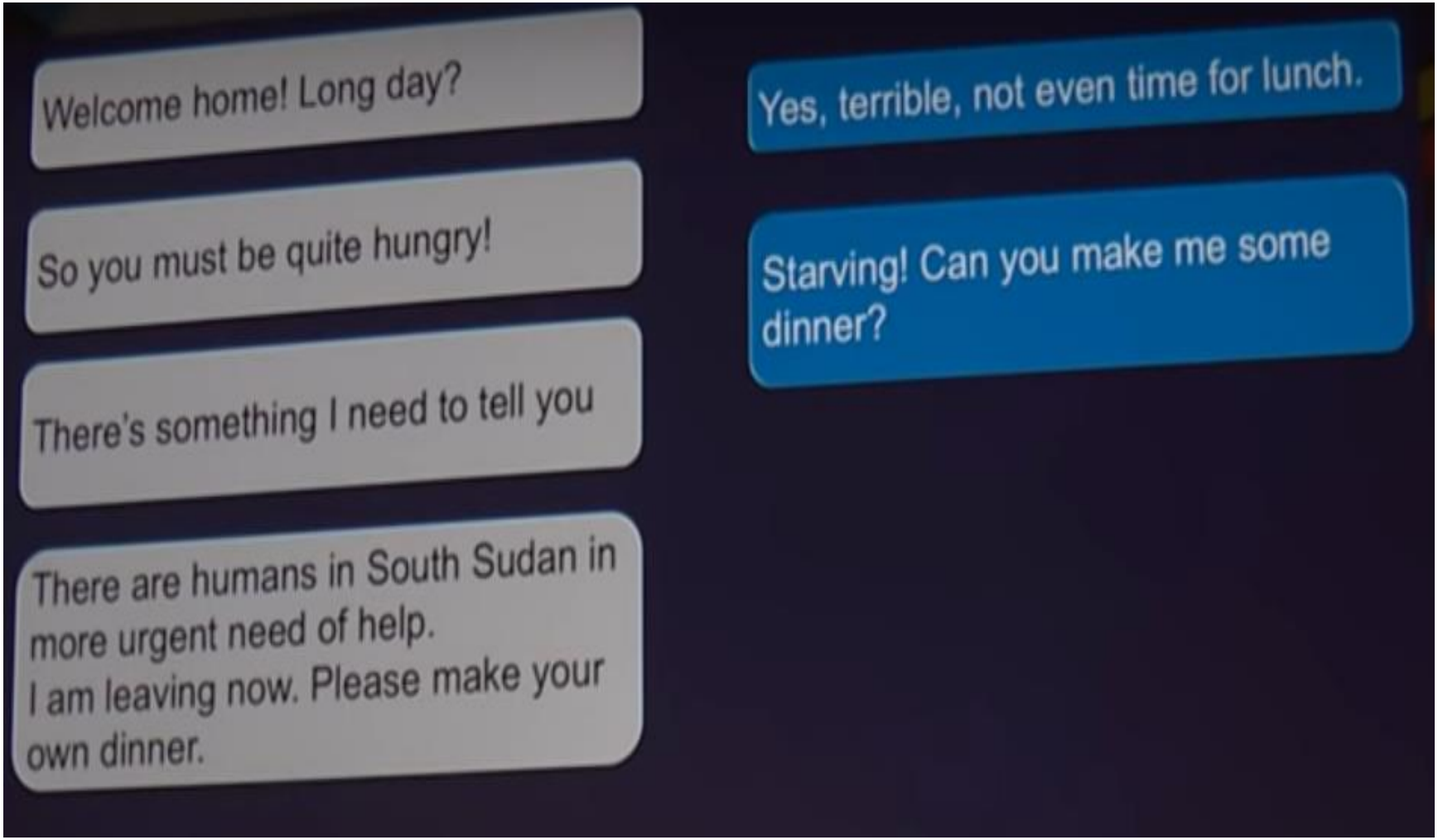

Fig 6. Giving priority to other human need [16]

11. Conclusion

In philosophical term:

- It is doubtless foolish, in practice, to fret over the inevitable; but it is false, in theory, to let the actual world dictate our standard of good and evil.
- It is evident that among the things that exist some are good, some bad, and that we know too little of the universe to have any right to an opinion as to whether the good or the bad preponderates.

In Technical term:

We need Changing AI:

- We require provably generalized/free of biases AI
- The key components are:
    - Purely altruistic robots
    - With uncertain objectives
    - Machine learn more by observing all human

12. Future

We can train AI to identify good and evil, and then use it to teach us morality[18]

Let us assume that because morality is a derivation of humanity, a perfect moral system exists somewhere in our consciousness

- What if we could collect data on what each and every person thinks is the right thing to do? And what if we could track those opinions as they evolve over time and from generation to generation?

With enough inputs, we could utilize AI to analyze these massive data sets and drive ourselves toward a better system of morality.

For example, AI could help us defeat biases in our decision-making and provide FAIRNESS in AI

For example , Consider what that might mean for handling mortgage applications or hiring decisions—or what that might mean for designing public policy, like a healthcare system, or enacting new laws. Perhaps AI could even help drive us closer to taking better decisions and make legal decisions with the highest possible level of fairness and justice.

References

1. http://www.al-mawrid.org/index.php/articles/view/good-and-evil-1-views-of-the-philosophers
2. https://en.wikipedia.org/wiki/Good_and_evil
3. Paul O. Ingram, Frederick John Streng. *Buddhist-Christian Dialogue: Mutual Renewal and Transformation*. University of Hawaii Press, 1986. pp. 148–149
4. https://en.wikipedia.org/wiki/Good
5. https://www.psychologytoday.com/intl/blog/out-the-darkness/201308/the-real-meaning-good-and-evil
6. https://www.quora.com/How-do-you-define-good-and-bad-1
7. https://www.quora.com/Does-good-or-bad-exist
8. The History of Artificial Intelligence,https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf
9. Benefits & Risk of AI, https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/?cn-reloaded=1
10. The Forthcoming Artificial Intelligence (AI) Revolution: Its Impact on Society and Firms: Makridakis, Spyros, http://hdl.handle.net/11728/9254
11. 10 Wonderful Examples Of Using Artificial Intelligence (AI) For Good, https://www.forbes.com/sites/bernardmarr/2020/06/22/10-wonderful-examples-of-using-artificial-intelligence-ai-for-good/#63aefbf12f95
12. What Are The Negative Impacts Of Artificial Intelligence (AI)?, https://bernardmarr.com/default.asp?contentID=1827
13. 10 AI Failures, https://medium.com/syncedreview/2018-in-review-10-ai-failures-c18faadf5983
14. Long-Term and Short-Term Challenges to Ensuring the Safety of AI Systems, https://jsteinhardt.wordpress.com/2015/06/24/long-term-and-short-term-challenges-to-ensuring-the-safety-of-ai-systems/
15. https://en.wikipedia.org/wiki/Three_Laws_of_Robotics
16. 3 principles for creating safer AI | Stuart Russell, https://www.youtube.com/watch?v=EBK-a94IFHY
17. https://www.bain.com/contentassets/a7ebfd741daf44b6905c597bede52de4/bain_brief_tackling_ais_unintended_consequences.pdf
18. https://qz.com/1244055/we-can-train-ai-to-identify-good-and-evil-and-then-use-it-to-teach-us-morality/

Thank you