AI ethics

Tae Wan Kim
Associate Professor of Business Ethics
Tepper School of Business
1. Coexistence?
The future trolley problem

- **Sentience (Utilitarianism):** The capacity for phenomenal experience or quaila, such as the capacity to feel pain and suffer.

- **Sapience (Deontology):** A set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsible agent.
Qualia

Feelings and experiences vary widely. For example, I run my fingers over sandpaper, smell a skunk, feel a sharp pain in my finger, seem to see bright purple, become extremely angry. In each of these cases, I am the subject of a mental state with a very distinctive subjective character. There is something it is like for me to undergo each state, some phenomenology that it has. Philosophers often use the term ‘qualia’ (singular ‘quale’) to refer to the introspectively accessible, phenomenal aspects of our mental lives. In this broad sense of the term, it is difficult to deny that there are qualia. Disagreement typically centers on which mental states have qualia, whether qualia are intrinsic qualities of their bearers, and how qualia relate to the physical world both inside and outside the head. The status of qualia is hotly debated in philosophy largely because it is central to a proper understanding of the nature of consciousness. Qualia are at the very heart of the mind-body problem.

The entry that follows is divided into ten sections. The first distinguishes various uses of the term ‘qualia’. The second addresses the question of which mental states have qualia. The third section brings out some of the main arguments for the view that qualia are irreducible and non-physical. The remaining sections focus on functionalism and qualia, the explanatory gap, qualia and introspection, representational theories of qualia, qualia as intrinsic, nonrepresentational properties, relational theories of qualia and finally the issue of qualia and simple minds.

1. Uses of the Term ‘Qualia’
2. Which Mental States Possess Qualia?
3. Are Qualia Irreducible, Non-Physical Entities?
4. Functionalism and Qualia
5. Qualia and the Explanatory Gap
6. Qualia and Introspection
7. Representational Theories of Qualia
Abstract

We attempt to determine the discriminability and organization of neural activation corresponding to the experience of specific emotions. Method actors were asked to self-induce nine emotional states (anger, disgust, envy, fear, happiness, lust, pride, sadness, and shame) while in an fMRI scanner. Using a Gaussian Naive Bayes pooled variance classifier, we demonstrate the ability to identify specific emotions experienced by an individual at well over chance accuracy on the basis of: 1) neural activation of the same individual in other trials, 2) neural activation of other individuals who experienced similar trials, and 3) neural activation of the same individual to a qualitatively different type of emotion induction. Factor analysis identified valence, arousal, sociality, and lust as dimensions underlying the activation patterns. These results suggest a structure for neural representations of emotion and inform theories of emotional processing.
Carnegie Mellon University Drama

Training data

Deep Learning Neural Network

SAD

Sad
Happy Shame
The future trolley problem

- **Sentience (Utilitarianism):** The capacity for phenomenal experience or *quaila*, such as the capacity to feel pain and suffer.

- **Sapience (Deontology):** A set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsible agent.
2. AI as scapegoat
Consider the following scenario: You are a person from a racially underrepresented group, say, X, and you recently applied to an online mortgage approval system and were rejected. The bank that hosts the online application system has recently started using AI to recommend mortgage applications for approval. You happened to know that the bank’s approval rate for clients of the same race as yours has recently abnormally decreased, for no good reason. You meet with a representative of the bank and claim that the bank has racially discriminated against you, and that the bank should be held liable for the discrimination. The bank representative says that it is impossible for the autonomous artificial agent to discriminate racially against applicants, because the algorithms it uses were designed to be indifferent to the race of applicants. To prove that, in front of you the representative submits ten fake applications equally qualified as yours (as judged by independent human evaluators) that consist of 5 whites and 5 Xs. The AI accepts all white applicants but only 2 Xs. The representative looks puzzled.
The Scapegoat Argument

P1) “Agent A is responsible for Act X” means just that X is properly *attributable* to A in a way that renders A open to moral appraisal for performing X.

P2) Agent A is open to moral appraisal for Act X just when X is expressive of A’s reflective or deep self or practical agency.

P3) Action X is expressive of Agent A’s self or agency only when X identifies with A’s desires, reasons, attitudes, or commitments that move A to perform X (whereas X is not expressive of A’s self or agency when X does not identify A’s desires, reasons, attitudes, or commitments, especially when A does not have volition or control over doing X or A cannot be aware of X).

P4) In the mortgage bank case, the racial discrimination was not expressive of any humans’ desires, reasons, attitudes, or commitments, and none of the humans’ practical identities moved the thinking machine to racially discriminate. (The humans did not have volition or control over the autonomous artificial mortgage appraiser’s creating the emergent property of racial discrimination and the humans in the bank were not able to be aware of the autonomous machine’s discriminative appraisal)

C) Thus, the humans in the bank are not responsible for the outcome action.
Principle of Fair Reciprocity

• If accidental or unforeseeable harm is an inevitable externality of freedom of action, a just society should implement a reasonable principle to fairly allocate the cost of unforeseeable harms.

• In a liberal society in which equal and free persons, who have different conceptions of good, live together, reciprocity is one of the few agreed upon principles. Reciprocity here means that **burdens must be borne by benefits.**

• The cost of unforeseeable harms created by a company that uses AI must be proportionately aligned with the benefit that companies and other parties gain by using AI.

• One efficient way to require companies that use AI to take the proportionate responsibility to remedy unforeseeable harms. By doing so, the burden is accordingly apportioned across companies and across customers who benefit from the companies’ AI services.
3. Super-intelligence and “Existential risk”

<table>
<thead>
<tr>
<th>Scope</th>
<th>Thinning of the ozone layer</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>global</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>local</td>
<td>Recession in a country</td>
<td>Genocide</td>
</tr>
<tr>
<td>personal</td>
<td>Your car is stolen</td>
<td>Death</td>
</tr>
<tr>
<td></td>
<td>endurable</td>
<td>terminal</td>
</tr>
</tbody>
</table>

*Figure 1. Six risk categories*
Proportion of experts with 10% 50% 90% confidence of HLMI by that date
## From HLAI to Superintelligence

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within 2 years</td>
<td>10%</td>
<td>19%</td>
<td>24</td>
</tr>
<tr>
<td>Within 30 years</td>
<td>75%</td>
<td>62%</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>%</th>
<th>PT-AI</th>
<th>AGI</th>
<th>EETN</th>
<th>TOP100</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely good</td>
<td>17</td>
<td>28</td>
<td>31</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>On balance good</td>
<td>24</td>
<td>25</td>
<td>30</td>
<td>40</td>
<td>28</td>
</tr>
<tr>
<td>More or less neutral</td>
<td>23</td>
<td>12</td>
<td>20</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>On balance bad</td>
<td>17</td>
<td>12</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Extremely bad (existential catastrophe)</td>
<td>18</td>
<td>24</td>
<td>6</td>
<td>8</td>
<td>18</td>
</tr>
</tbody>
</table>
The good-story bias

• “Our intuitions about which future scenarios are plausible and realistic are shaped by what we see on TV and in movies and what we read in novels…..We should then suspect our intuitions of being biased in the direction of overestimating the probability of those scenarios that make for a good story, since such scenarios will see much more familiar and more real.” Nick Bostrom