Centre for Educational Research and Innovation — CERI

Assessing AI Capabilities for Policymakers

Stuart W. Elliott, AI and the Future of Skills Project
OECD, Education and Skills Directorate

IJCAI Evaluation Beyond Metrics Workshop
Vienna, 24 July 2022

Team: Abel Baret, Margarita Kalamova, Aurelija Masiulyte, Nóra Réva, Mila Staneva

OECD
BETTER POLICIES FOR BETTER LIVES

Session Overview

- OECD project to assess AI capabilities for policymakers (0:15)
  - Project goal
  - Desirable characteristics for the AI measures and some implications
  - Initial project directions
  - Need for a conceptual framework: questions to the field

- Panel responses (0:15)
  - Virginia Dignum, Umeå University
  - Tony Cohn, University of Leeds
  - Songül Tolan, Joint Research Center, European Commission

- Discussion (0:15)

OECD Project: AI and the Future of Skills (AIFS)

- Project goal: Develop measures of AI capabilities for policymakers
  - Communicate AI strengths and limitations that have implications for work and education
    - What human capabilities will be too difficult for AI and robotics to reproduce over the next few decades?
    - What education and training will be needed to allow most people to develop capabilities that are beyond the capabilities of AI and robotics?
  - Track changes in AI strengths and limitations over time

Economic Indicators as an Analogy

- Point of the analogy: NOT your area of expertise
- Key economic indicators for policymakers and the public
  - Economic output (gross domestic product/GDP)
  - Inflation
  - Unemployment
- Broadly familiar to non-experts
  - Provide motivation for important policies
  - Seem “obvious” though they required two centuries of development
  - Good enough though still conceptually flawed
- Premise of AIFS: We need measures like this for AI capabilities

Desirable Characteristics for AI Measures for Policy

Not just valid and reliable, but also:

- Understandable
- Comprehensive
- Repeatable
- Policy Relevant

Desirable Characteristics: Understandable

- Implies: Small number of measures with meaningful scales
- Example: OECD test of adult competences
  - Three measures: literacy, numeracy, problem solving
  - 5-level scale for literacy
    - Level 1: short texts, locating single piece of information, no competing information, basic vocabulary
    - Level 3: long texts, interpret information with inference, multistep operations, need to disregard irrelevant information
    - Level 5: integrate information over several long texts, synthesize and contrast ideas, evaluate and choose evidence, use subtle rhetorical information

Desirable Characteristics: Comprehensive

- Implies: One measure for each major aspect of AI capability
  - Include everything policymakers should understand
- Severe constraint if only a small number of measures
- How to choose "major aspects" of AI capability?
  - Perhaps: Different capabilities
    - Language, problem solving, sensory perception, motor control, social interaction … [Cf. Hernández-Orallo and Vold, 2019]
  - Perhaps: Underlying cognitive processes
    - Learning, knowledge representation, reasoning, integration, … [Cf. Forbus, 2021]

Desirable Characteristics: Repeatable

- Implies: Inexpensive and fast to produce the measures
- This is relative: economic and education indicators aren’t cheap
  - Probably >€100 million per “round”
- Start low and build up to conceptually adequate approaches
  - Initially:
    - Expert judgement
    - Using existing measures
  - Then argue for what would be needed to do it right

Desirable Characteristics: Policy Relevant

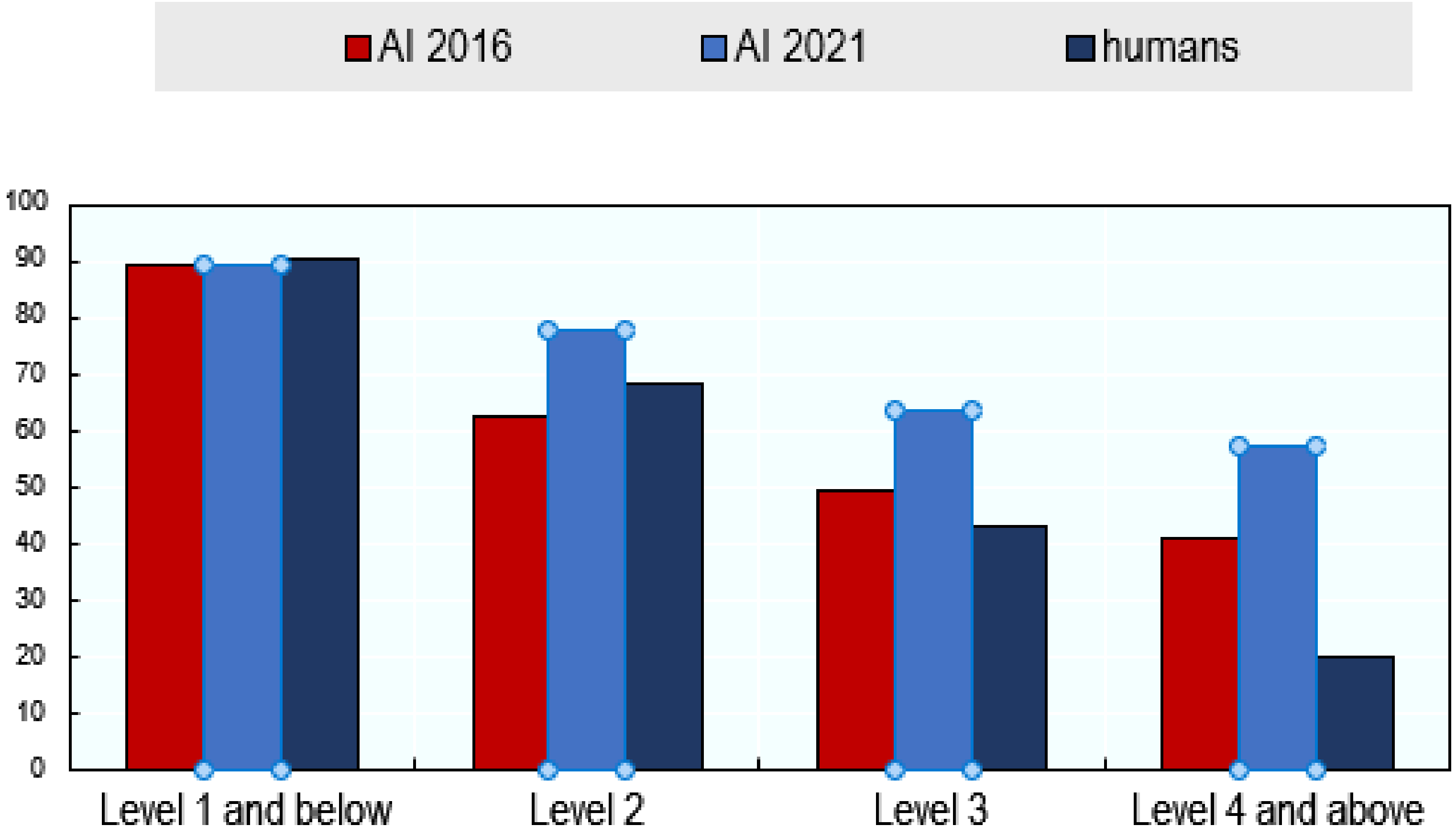
- Implies: Possible to link to education and work
- AIFS policy questions
  - What human capabilities will be too difficult for AI and robotics to reproduce over the next few decades?
  - What education and training will be needed to allow most people to develop capabilities that are beyond the capabilities of AI and robotics?
- Links to education and work are complex
  - Better to think of measures that can support understanding of AI that can be used to consider a variety of implications: educational attainment, curriculum redesign, jobs threatened by automation, job redesign, …

AIFS Initial Directions

- Exploring three types of tests
  - Educational tests – for link to formal education
  - Occupational tests – for link to work tasks
  - AI benchmark tests and evaluations
- Approaches
  - Educational and occupational tests
    - Expert judgement of individual test questions
    - Considering value of commissioning AI systems
  - AI benchmark tests and evaluations
    - Analysis and synthesis of existing measures
    - Considering value of commission new measures

AIFS Initial Directions: Example Educational Test

Expert judgement of AI capabilities for tasks on OECD adult literacy test



AIFS Initial Directions: AI benchmark tests/evaluations

- Analysis and synthesis work
  - Tony Cohn and José Hernández-Orallo are developing descriptors of evaluation instruments from AI that could potentially be used to identify better measures
  - Guillaume Avrin, Swen Ribeiro, and Elena Messina are summarising evaluations of AI and robotics tasks by the Laboratoire national de métrologie et d'essais (LNE) and the US National Institute of Standards and Technology (NIST)
  - Yvette Graham is developing an integrated measure of natural language capabilities based on a set of available AI benchmark tests
  - Potential additional work to connect detailed task descriptions to existing AI benchmarks
  - Potential additional work to develop tasks to assess aspects of AI capabilities that are clearly beyond current techniques

Conceptual framework to structure the AIFS measures

- Assessment 101: what are we are trying to measure?
- Questions for the field
  - What are the key strengths and limitations that an evaluation of AI capabilities should measure to help policymakers understand and track development of AI over the next several decades?
    - Especially: current limitations where a breakthrough could have big implications
  - What aspects of these strengths and limitations should be measured by direct evaluation instruments and which would be better measured by expert judgement?

Thank you

We’re looking for computer and cognitive scientists who would like to be involved!

For more information:
https://www.oecd.org/education/ceri/future-of-skills.htm
futureofskills@oecd.org or stuart.elliott@oecd.org