SUPERINTELLIGENCE

TWO APPROACHES FOR AI

- TASK BASED APPROACH: DESIGN SPECIFIC WAY TO APPLY AI TO EACH TASK
- GENERALIZED-BASED APPROACH: UNDERSTAND NEW TASKS WITH LITTLE OR NO TASK-SPECIFIC TRAINING BY GENERALIZING FROM PREVIOUS EXPERIENCE (BOTTOM-UP) ←

POTENTIAL OF THE GENERALIZATION-BASED APPROACH IN HOW HUMANS DEVELOPED

- SKILL OF ABSTRACTION: EXTRACT COMMON STRUCTURE FROM DIFFERENT SITUATIONS = MORE EFFICIENT UNDERSTANDING
- COMMUNICATION SKILLS-THEORIES: SHARE OUR IDEAS

HUMANS HAVE LIMITATIONS THAT AGI NOT HAVE

- SPEED DIFFERENCE BETWEEN NEURONS-TRANSISTORS
- BRAIN SIZE
- FASTER THAN EVOLUTION

↑

→ MOST LIKELY WAY TO ACHIEVE IS THROUGH A COLLECTIVE AGI COMPOSED OF MULTIPLE AGIs

REACHING SUPERINTELLIGENCE → "INTELLECT THAT GREADLY EXCEEDS THE COGNITIVE PERFORMANCE IN ALL DOMAINS OF INTEREST, BETTER THEN ALL OF HUMANITY COORDINATING GLOBALLY" ← BOSTROM

↓

EXEED HUMAN PERFORMANCE IN TERMS OF PROCESSING SPEED AND SIZE OF NEURAL NETWORK

+

POTENTIAL OF REPLICATION, CULTURAL LEARNING AND RECURSIVE IMPROVEMENT ← WILL BECOME GREATER AS AIs BECOME MORE INTELLIGENT

3 FACTOR THAT WILL BECOME MORE IMPORTANT AS AIs BECOME MORE INTELLIGENT

- REPLICATION:
  - AIs LESS CONSTRAINED THAN HUMANS. EASY TO CREATE A DOUPLICATE OF IT
  - AGIS DECOMPOSE DIFFICULT TASKS IN SUBTASKS.
  - COLLECTIVE SUPERINTELLIGENCE: LARGE GROUP OF AGIs TO CARRY OUT MORE COMPLEX TASKS
- CULTURAL LEARNING:
  - SKILL TO ACQUIRE/SHARE KNOWLEDGE. COLLECTIVELY AGI SOLVE HARDER PROBLEM THAN INDIVIDUAL AGIS
- RECURSIVE IMPROVEMENT:
  - IMPROVE TRAINING PROCESSES TO DEVELOP THEIR SUCCESSOR

GOAL AND AGENCY

AGENCY: ABILITY OF AN AGENT TO HAVE ITS OWN GOALS:

- DESIGNED OBJECTIVES: GOALS THAT AN AI HAS BEEN DESIGNED TO ACHIEVE
- OWN GOALS: GOALS THAT AN AI WANTS TO ACHIEVE

- CURRENT AI SYSTEMS: ACHIEVE DESIGN OBJECTIVES WITHOUT TRULY UNDERSTANDING WHAT ARE/ACTIONS TO ACHIEVE THEM

- BOUNDED RATIONALITY: SYSTEM CAN TRY TO ACHIEVE A ROLE W/OUT TAKING THE BEST ACTIONS.

3 WAY HOW AN AI COULD GAIN POWER

- AIs PURSUE POWER AS AN INSTRUMENTAL GOAL TO ACHIEVE OTHER GOALS] THERE ARE INSTRUMENTAL GOAL THAT INCREASE THE CHANCES OF AN AGENT'S FINAL GOALS BEING REACHED
  - SELF-PRESERVATION
  - RESOURCE ACQUISITION
  - SELF-IMPROVEMENT
- AIs PURSUE POWER FOR ITS OWN SAKE
- AIs GAIN POWER WITHOUT AIMING FOR IT

GOAL-DIRECTED

1) SELF-AWARENESS: UNDERSTAND THAT AN AGENT IS PART OF THE WORLD AND THAT ACTIONS HAVE CONSEQUENCES
2) PLANNING: MAKE DECISIONS WRT THE OUTCOME VALUE
3) CONSEQUENTIALISM: DECIDE BEST PLANS WRT CONSEQUENCES OF ACTION
4) SCALE: TAKE INTO ACCOUNT EFFECTS OF PLANS
5) COHERENCE: REMAIN INTERNALLY UNIFIED IN IMPLEMENTING THE SIGNLE BEST ACTION
6) FLEXIBICITY: ADAPT PLANS FLEXIBICY

→ WHAT IS MISSING

1) TRAINED ON 3RD-PERSON DATA, NO 1ST-PERSON PERSPECTIVE
2) LIMITED TYPES OF PLAN CONSIDERED. NOT ALWAYS EXTENSIVE PLANS
3) SUBJECTIVE PREFERENCES ABOUT ACTIONS WRT CONSEQUENCES
4) TRAIN IN SMALL SCALE ENVIORMENT/NO GENERALIZATION ABILITY
5) INTERNAL CONFLICT (LIKE HUMANS)
6) 1 INITIAL PLAN. NO ADAPTATION TO NEW SITUATIONS/RETHINK PLANS