TURING: COMPUTE MACHINERY AND INTELLIGENCE: REPLACE

↑ CRITIQUE

SEARLE: IS INADEQUATE. IMITATION NOT SUFFICENT FOR UNDERSTANDING + CHINESE ROOM (WEAK VS STRONG AI)

PSYCHOLOGICAL COGNITIVE ⇒ EMPIRICAL BEHAVIOURISTIC

ETHICS ← NORMS

- SYSTEMATIC REFLECTION ABOUT WHAT IS MORAL
- WHAT ONE SHOULD DO, ALL THINGS CONSIDERED

- DESCRIPTIVE ETHICS
- NORMATIVE ETHICS — FIRST - ORDER MORAL THEORY

FOUR THEORY ←

- DEONTOLOGY: WHAT ONE SHOULD DO? OUR DUTY
  - ONE OR MORE DUTY
  - EACH THEORY TELLS WHAT PRINCIPLE TO FOLLOW
  - MORAL DILEMMAS
  - HIERARCHY OF PRINCIPLES
  - * MONISM → KANT, GOLDEN RULE
  - * PLURALISM → MULTIPLE DUTIES
- VIRTUE ETHICS:

- META-ETHICS: SECOND - ORDER MORAL THEORY
  - SEMANTIC → MEANING OF GOOD RIGTH, CORRECT /
  - EPISTEMEOLOGICAL → HOW?
  - ONTOLOGICAL → DO SUCH VALUES EXISTS?
  - # EMOTIVISM → BOO ABORTION (STEVENSON)
  - * NON-COGNITIVISM → CONTEXT OF A WORD / ITS MOOD (HARE)
  - * EXPRESSIVISM → NOT OBJECTIVE TRUTHS BUT PERSONAL EVALUATIONS
- APPLIED ETHICS: THIRD -ORDER MORAL THEORY

TURING: COMPUTE MACHINERY AND INTELLIGENCE: REPLACE

PSYCHOLOGICAL COGNITIVE ⇒ EMPIRICAL BEHAVIORISTIC

↑ CRITIQUE

SEARLE: IS INADEQUATE. IMITATION NOT SUFFICENT FOR UNDERSTANDING + CHINESE ROOM (WEAK VS STRONG AI)

ETHICS ← NORMS

- SYSTEMATIC REFLECTION ABOUT WHAT IS MORAL
- WHAT ONE SHOULD DO, ALL THINGS CONSIDERED

FOUR THEORY

- DESCRIPTIVE ETHICS
- NORMATIVE ETHICS — FIRST-ORDER MORAL THEORY

- META-ETHICS: SECOND-ORDER MORAL THEORY
  - SEMANTIC → MEANING OF GOOD / RIGHT, CORRECT
  - EPISTEMEOLOGICAL → HOW? → COGNITIVISM → MORAL KNOWLEDGE = ALL OTHER K. / NON-COGNITIVISM ≠ (INVOLVE NON-COGNITIVE MENTAL STATES)
  - ONTOLOGICAL → DO SUCH VALUES EXISTS? → REALISM → MORAL FACT AND VALUES ∃ / ANTI-REALISM → ∄
  - # EMOTIVISM → BOO ABORTION (STEVENSON)
  - # NON-COGNITIVISM → CONTEXT OF A WORD → ITS MOOD (HARE)
  - # EXPRESSIVISM → NOT OBJECTIVE TRUTHS BUT PERSONAL EVALUATIONS
- APPLIED ETHICS: THIRD-ORDER MORAL THEORY

MORAL UNCERTAINTY: WHICH FIRST-ORDER MORAL THEORY IS CORRECT?

ETHICAL ARTIFICIAL AGENTS: WHICH THEORY?

- MININTERPRET THE RULES
- LEAD UNFORESEEN SCENARIOS
- PROBLEMS W/ HUMAN ONTOLOGY ALREADY

- USE ETHICAL DILEMMAS TO TRAIN
- ASSUMPTION IN TRAINING: ARE THERE CORRECT ANSWERS TO TRAIN?

- SOME ELEMENTS OF ML RESAMBLE THE THEORY
- TO WHAT EXTEND CLOSE ENOUGH TO PREDICT THE CONSEQUENCES OF YOUR ACTION?
- HOW TO EVALUATE CONSEQUENCES?

COMBINED APPROACH →

- DEONTOLOGY + CONSEQUENTIALISM

UTILITARIANISM UNTIL SACRED VALUES. → USE DEONTOLOGICAL

COMBINED APPROACH →

- HYBRID APPROACH
  - TOP-DOWN: THEORY-DRIVEN REASONING
  - BOTTOM-UP: REASONING SHAPED BY EVOLUTION AND LEARNING

- DEONTOLOGY: WHAT ONE SHOULD DO? OUR DUTY
  - ONE OR MORE DUTY
  - EACH THEORY TELLS WHAT PRINCIPLE TO FOLLOW
  - MORAL DILEMMAS
  - HIERARCHY OF PRINCIPLES
  - # MONISM → KANT, GOLDEN RULE
  - # PLURALISM → MULTIPLE DUTIES
- VIRTUE ETHICS: → DO WHAT IS VIRTUOUS + GOOD SOLDIER
- PARTICULARISM: → CONTEXT-TO-CONTEXT
  - NO GENERAL PRINCIPLE OF ACTION
  - WHAT IS FAVORABLE, WHAT AGAINST (CAN CHANGE DEPENDING ON THE SITUATION)
- CONSEQUENTIALISM: → WHAT TO DO DEPENDS ON THE CONSEQUENCES OF YOUR ACTION
  - # HEDONISM
  - # UTILITARIANISM
    - ACT UTILITARIANISM
    - RULE UTILITARIANISM
  - # EGOISM

SOCIETAL PROBLEM: NO AGREEMENT ON WHICH THEORY IS THE CORRECT ONE + DIFFERENT THEORIES NOT ALWAYS COMPATIBLE

SUPERINTELLIGENCE

HUMANS HAVE LIMITATIONS THAT AGI NOT HAVE
- SPEED DIFFERENCE BETWEEN NEURONS-TRANSISTORS
- BRAIN SIZE
- FASTER THAN EVOLUTION

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

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

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

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

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 SUB TASKS.
  - 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.

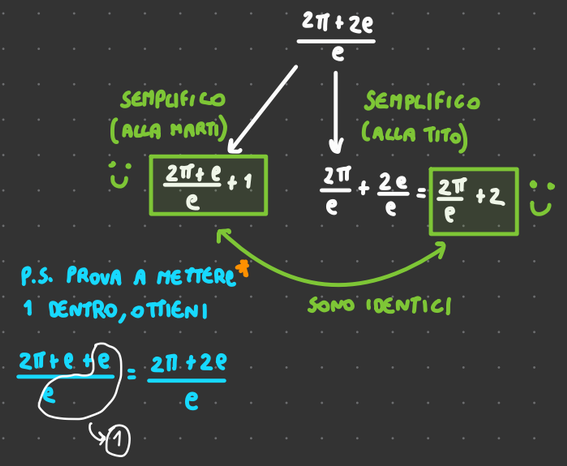
3 WAY HOW AN AI COULD GAIN POWER
- AIS PURSUE POWER AS AN INSTRUMENTAL GOAL TO ACHIEVE OTHER GOALS
- AIS PURSUE POWER FOR ITS OWN SAKE
- AIS GAIN POWER WITHOUT AIMING FOR IT

THERE ARE INSTRUMENTAL GOAL THAT INCREASE THE CHANCES OF AN AGENT'S FINAL GOALS BEING REACHED
- SELF-PRESERVATION
- RESOURCE ACQUISITION
- SELF-IMPROVEMENT

GOAL-DIRECTED → WHAT IS MISSING

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 FLEXIBLY

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 ABICITY
5) INTERNAL CONFLICT (LIKE HUMANS)
6) 1 INITIAL PLAN. NO ADAPTATION TO NEW SITUATIONS/RETHINK PLANS



* STO CONTROLLANDO SE E' GIUSTO