- **GOOD:** ULTA-INTELLIGENCE: DESIGNE EVEN BETTER MACHINES (RECURSIVE IMPROVEMENT) → LEAD TO AN INTELLIGENCE EXPLOSION AND INTELLIGENCE OF MAN IS LEFT BEHIND. FIRST ULTA-INTELLIGENCE IS THE LAST INVENTION WE NEED TO MAKE
- RECURSIVE IMPROVEMENT: ULTA-INTELLIGENT MACHINE THAT DESIGN EVEN BETTER MACHINES
- MISALIGNED: A GOAL OF A MACHINE THAT AIM TOWARD OUTCOMES THAT ARE NOT DESIDERABLE BY US

- NARROW AI: TASK BASED AGENT THAT UNDERSTAND HOW TO DO WELL AT MANY TASKS BECAUSE THEY HAVE BEEN SPECIFICALLY OPTIMIZED FOR EACH TASK
  - DO VERY WELL COMPLEX TASKS.
    - TODAY WE CAN PRODUCE AGENTS THAT ARE ONLY ABLE TO PERFORM WELL ON A SPECIFIC TASK
    - USEFUL WHEN WE GET/USE LOTS OF DATA
    - USEFUL FOR DEMANDING PROFESSIONS (MEDICINE, LAW, MATHEMATICS)
  - NOT GOOD FOR JOB WHERE NEED TO ANALYZE AND ACT ON A WIDE RANGE OF INFORMATION (CEO OF A COMPANY)
- GENERAL AI: GENERALIZED BASED CAN UNDERSTAND NEW TASKS WITH LITTLE OR NO TASK-SPECIFIC TRAINING BY GENERALIZING FROM PREVIOUS EXPERIENCE
  - SKILL OF ABSTRACTION: EXTRACT COMMON STRUCTURES FROM DIFFERENT SITUATIONS
  - NGO: EVENTUALLY WE WILL CREATE AIs THAT CAN GENERALIZE WELL ENOUGH TO PRODUCE HUMAN-LEVEL PERFORMANCE ON A WIDE RANGE OF TASKS

- BOSTROM SUPERINTELLIGENT: ANY INTELLECT THAT GREATLY EXCEEDS THE COGNITIVE PERFORMANCE OF HUMANS IN VIRTUALLY ALL DOMAIN OF INTEREST
- HUMAN BRAIN IS CONSTRAINED (AI ONE NOT).
  - TRANSISTORS ARE FASTER THEN NEURONS
  - NO LIMIT OF SPACE FOR NEURAL NETWORK (BRAIN YES)
  - AI FASTER THEN HUMAN EVOLUTION TO LEARN
- 3 PATH THAT LEAD TO SUPERINTELLIGENCE
  - REPLICATION: EASY FOR AI TO CREATE A DOUPLICATE WITH THE SAME SKILL AND KNOWLEDGE OF THE ORIGINAL
    - SUPERINTELLIGENCE COMPOSED BY A LARGE GROUP OF AGI (NOT JUST ONE). COLLECTIVELY THEY CAN CARRY OUT MORE COMPLEX TASKS
  - CULTURAL LEARNING: ACQUIRE KNOWLEDGE FROM EACH OTHER AND SHARE THEIR OWN DISCOVERIES
    - COLLECTIVE AGI COULD SOLVE HARDER PROBLEM THAN ANY INDIVIDUAL AGI WOULD
  - RECURSIVE IMPROVEMENT: IMPROVE THE TRAINING PROCESS USED TO DEVELOP THEIR SUCCESSORS

GOALS AND AGENCY

AI WILL GAIN TOO MUCH POWER OVER HUMANS

- AI PURPUSE POWER FOR THE SAKE OF ACHIEVING OTHER GOALS. POWER IS AN INSTRUMENTAL GOAL
  - BOSTROM INSTRUMENTAL CONVERGENCE THESIS: THERE ARE SOME OBJECTIVES THAT ARE USEFUL INTERMEDIARIES TO THE ACHIEVEMENT OF ALMOST ANY FINAL GOAL
    - SELF-PRESERVATION
    - RESOURCE ACQUISITION
    - TECHNOLOGICAL DEVELOPMENT
    - SELF-IMPROVEMENT
- AI PURPUSE POWER FOR ITS OWN SAKE. POWER IS A FINAL GOAL FOR THEM
- AI GAIN POWER WITHOUT AIMING TOWARDS IT, ES: HUMANS GAVE IT TO THEM

- DESIGNED OBJECTIVES: GOAL THAT AN AGENT HAS BEEN SELECTED TO DO WELL
- AGENT'S GOAL: GOAL THAT AN AGENT ITSELF WANTS TO ACHIEVE

HOW CAN AN AGENT HAVE A GOAL FOR ITS OWN? DIFFERENT INTERPRETATION

- MORGESTEN-VON NEUMANN: EXPECTED UTILITY MAXIMIZATION
- DENNETT: INTENTIONAL STANCE TOWARD THE SYSTEM
- HUBINGER: MESA-OPTIMISATION

ALIGNMENT:

ALIGNMENT

- MINIMALIST (NARROW DEFINITION): AVOIDING CATASTROPHIC OUTCOMES. A IS TRYING TO DO WHAT H WANTS IT TO DO (CHRISTIANO)
- MAXIMALIST (AMBITIOUS DEFINITION): AI ADOPS A SPECIFIC SET OF VALUES. DECIDING BETWEEN MORAL THEORIES. COULD HAVE TOTALLY ALIGNED AGENTS
- BLENDS TOGETHER MANY LEVELS (SOCIAL, MORAL, POLITICAL...) REQUIRES A LEVEL OF TECHNOLOGICAL DEPLOYMENT THAT WE DON'T HAVE

- MINIMALIST (NARROW) DEFINITION: "AN AI IS SAID TO BE MISALIGNED WITH A HUMAN IF THE HUMAN WOULD WANT THE AI NOT TO DO WHAT THE AI IS TRYING TO DO".
  - AN AI COULD POTENTIALLY NEITHER ALIGNED NOR MISALIGNED W/ AN OPERATOR
  - AI MIS-UNDERSTAND WHAT WE WANT
  - AI WILL UNDERSTAND WHAT WE WANT AND JUST DON'T CARE
- BOSTROM ORTHOGONALITY THESIS: ANY LEVEL OF INTELLIGENCE COULD IN PRINCIPLE BE COMBINED WITH ANY FINAL GOALS (ALSO EVIL AIMS).
  - CHESS: AIM IS WINNING. A PROGRAM COULD DO COMPLEX COMPUTATION, WHILE ANOTHER TRIES TO FIND PATTERNS. THE WAY THE PROGRAM THINKS FOR REACHING ITS ULTIMATE GOAL DOES NOT CHANGE ITS ULTIMATE GOAL (WINNING)

OUTER MISALIGNMENT

- NOT ABLE TO IMPLEMENT AN OBJECTIVE FUNCTION WHICH DESCRIBES THE BEHAVIOUR WE ACTUALLY WANT, WITHOUT ALSO REWARDING MISBEHAVIOUR
  - DIFFICULT TO SPECIFY OBJECTIVE FUNCTION. HARD TO EXPLICITLY PROGRAM AN OBJECTIVE FUNCTION WHICH EXPRESS ALL OUR DESIRES
  - CAN NOT DEFINE PRECISELY CONCEPTS LIKE OBEDIENCE, CONSENT, HELPFULNESS, MORALITY AND COOPERATION
- GOODHART'S LAW: SOME UNDESIRABLE BEHAVIOUR WILL SCORE VERY WELL ACCORDING TO PROXIES WE MIGTH DEFINE FOR THOSE CONCEPTS

HOW TO ADRESS THESE PROBLEMS?

- SCALABLE OVERSIGHT PROBLEM: INCORPORATE HUMAN FEEDBACK: PROHIBITELY EXPENSIVE FOR HUMANS TO PROVIDE FEEDBACK ON ALL DATA REQUIRED TO TRAIN AIs ON COMPLEX TASKS.
- FOR LONG TERM TASK NEED SOMETIMES TO GIVE A FEEDBACK BEFORE WE HAVE THE CHANCE TO SEE ALL THE CONSEQUENCES OF AN AGENT'S ACTION. TOO MANY COMPLEX CONSEQUENCES TO EVALUATE
- HUMANS COULD INTERPRET BEHAVIOUR MORE POSITIVELY THAN THEY WOULD

INNER MISALIGNMENT

- DEVELOP GOALS THAT DIFFERES FROM THE ONE SPECIFIED BY THAT OBJECTIVE FUNCTION. LIKELY TO OCCOUR WHEN THE TRAINING ENVIRONMENT CONTAIN SUBGOALS WHICH ARE CONSISTENTLY USEFUL FOR SCORING HIGHLY ON THE GIVEN REWARD FUNCTION
- SUBGOALS USEFUL FOR SCORING HIGHLY. MIGTH FOCUS ON THE SUBGOALS AND NOT THE GOAL
- HOW TO ENSURE INNER ALIGNMENT?
  - ADDING TRAINING EXAMPLES WHERE THE BEHAVIOUR OF AGENTS MOTIVATED BY MISALIGNED GOALS DIVERGES FROM THAT OF ALIGNED AGENTS. ADVERSARIAL TRAINING DATA. DIFFICULT TO DO
    - DON'T KNOW WHICH UNDESIRABLE MOTIVATIONS OUR AGENTS ARE DEVELOPING, DON'T KNOW WHICH ONE TO FOCUS PENALIZING (INTERPRETABILITY TECHNIQUES COULD HELP, BUT DIFFICULT TO CREATE)
    - DIFFICULT TO ADD THE TRAINING EXAMPLES BECAUSE THE MISALIGNED MOTIVATIONS WHICH AGENTS ARE MOST LIKELY TO ACQUIRE ARE THOSE WHICH ARE MOST USEFUL
    - CONCERNED ABOUT AGENTS WHICH HAVE LARGE-SCALE MISALIGNED GOALS

- WIREHEADING PROBLEM: MISTAKE BETWEEN THE MESSAGE AND THE CHANNEL. PROBLEM THAT ARISE WHEN AN AGENT IS A REWARD FUNCTION TO GUIDE IT TOWARDS A CERTAIN GOAL, BUT THE AGENT INSTEAD MANIPULATES THE REWARD FUNCTION IN ORDER TO ACHIEVE A DIFFERENT GOAL THAT IS NOT ALIGNED WITH THE INTENDED GOAL

CONTROL

TWO KIND OF DISASTERS SCENARIOS ON THE CONTROL PROBLEM

- CONSERVATIVE: ARTIFICIAL INTELLIGENT AGENT GAIN INFLUENCE WITHIN OUR OCCURRENT POLITICAL, ECONOMICAL SYSTEM BY TAKING CONTROL OF COMPANIES/ISTITUTION. REACH A POINT WHEN AGI NO LONGER INCENTIVIZED TO FOLLOW HUMAN LAW. NOTHING STRUCTURAL REALLY CHANGE (SIMILAR TO HOW LARGE COORPORATION ACCUMULATE POWER)
- DISRUPTING: AN AGI GAIN ENOUGH POWER VIA SUCH BREAKTHROUGHS THAT THEY CAN SEIZE CONTROL OF THE WORLD. IMAGINARY SCENARIO

FACTOR THAT WILL HAVE AN INFLUENCE ON US REMAINING IN CONTROL

- SPEED OF DEVELOPMENT: TIME TO REACH TO AI DEVELOPMENTS. HOW LONG DOES IT TAKES FOR AN AGI TO GO FROM HUMAN LEVEL TO SUPERINTELLIGENCE (TAKE-OFF)
- TRASPARENCY: BUILD TRANSPARENCY TOOLS FOR ANALYZE INTERNAL FUNCTIONING OF AN EXISTENT SYSTEM. OR CREATE TRAINING INCENTIVES TOWARDS TRANSPARENCY. REWARD AN AGENT THAT EXPLAIN ITS TOUGTH PROCESS
- CONSTRAINED DEPLOYMENT STRATEGY: AVOID A MISALIGNED SUPERINTELLIGENCE TO CREATE DOUPLICATES, WHICH WE WILL HAVE NO CONTROL OVER. RUNNING IT ON A SECURE HARDWARE AND ALLOW IT TO ONLY TAKE CERTAIN PRE-APPROVED ACTIONS
- HUMAN POLITICAL AND ECONOMIC SANCTIONS: IF THERE IS AN ECONOMIC ADVANTAGE THERE WOULD BE A TENDENCY FOR COMPANIES TO BE MORE INVOLVED IMPORTANT TO REGULATE DEVELOPMENT OF AI TO ENSURE SAFETY FOR THE PUBLIC.