AI ETHICS

(agent with capacity to act)

- ETHICAL CHALLENGES POSED BY CURRENT AND NEAR-FUTURE AI
- NOT ONLY TECHNOLOGY, ALSO ABOUT WHAT HUMANS DO WITH IT
- ISSUE OF SUPERINTELLIGENT AI
- STUDY
  - MORAL BEHAVIOUR OF HUMANS: THEY DESIGN/MAKE/USE/TREAT AI SYSTEMS
  - BEHAVIOUR OF MACHINE: MACHINE ETHICS SINGULARITY DUE TO SUPERINTELLIGENT AI

ETHICS AND MORALITY

(often synonyms)

ETHICS

SYSTEMATIC REFLECTION ON WHAT IS MORAL

(analyze-reflect-evaluate on questions-arguments about moral choices that people make using instruments/notions)

REFLECTS ON QUESTIONS AND ARGUMENTS CONCERNING THE MORAL CHOICES PEOPLE CAN MAKE

(aim of an argument)

- JUSTIFY OR REFUSE A STATEMENT
- DEFENDING/ATTACKING A OPINION
- NECESSITY: IF PREMISES ARE TRUE ⇒ CONCLUSION TRUE
- ARGUMENTATION: HOW HUMAN SCANI DO/REACH CONCLUSION FROM PREMISISES

DESCRIPTIVE ETHICS

DESCRIBE EXISTING MORALITY (OPINION ABOUT GOOD AND EVIL...)

- DESCRIBE MORALITY AS A SET OF RULES THAT DETERMINES PEOPLE ACTION
- STUDY HOW WE ACTUALLY ACT

DESCRIPTIVE JUDGEMENT:
- DESCRIBE PRESENT, PAST, FUTURE MORALITY (WHAT IS/WAS/WILL BE)
- ARE TRUE OR FALSE

NORMATIVE ETHICS

JUDGES MORALITY AND TRIES TO FORMULATE NORMATIVE RACCOMODATION ABOUT HOW TO ACT OR LIVE

- GIVES INSTRUCTION/RECCOMMENDATIONS = GUIDE PEOPLES ACTION
- STUDY OF HOW WE HAVE TO ACT

NORMATIVE JUDGEMENT:
- INDICATES IF SOMETHING IS GOOD OR BAD
- HOW THE WORLD SHOULD BE VS HOW IT IS (is a prescription)

MORALITY

TOTALITY OF OPINIONS, DECISION AND ACTION WITH WHICH PEOPLE (individually and collectively) EXPRESS WHAT THEY THINK IS GOOD OR RIGTH

IS THE TOTALITY OF NORMS AND VALUES THAT ACTUALLY EXISTS IN SOCIETY

NORMS:
- WHAT CONCRETE ACTIONS ARE REQUIRED, FORBIDDEN OR PERMITTED
- HOW PEOPLE ARE SUPPOSED TO TREAT EACH OTHER
- MEANS TO REALIZE VALUES, CONCRETE RULES

VALUES:

LASTING CONVINCTIONS THAT PEOPLE FEEL SHOULD BE STRIVED FOR IN GENERAL (NOT JUST FOR THEMSELVES) FOR A GOOD LIFE/ RIGTH SOCIETY

- ABSTRACT = SET OF TYPES OF BEHAVIOUR, WHAT PEOPLE WISH TO ACHIEVE
- INTRINSIC VALUE: OBJECTIVE VALUE IN AND OF ITSELF
- INSTRUMENTAL VALUE: MEANS TO REALIZING AN INTRINSIC VALUE

AI RATIONALITY AND INTELLIGENCE

- WEAK AI: ALREADY TODAY PERFORM ONLY ON A SPECIFIC DOMAIN
- STRONG AI: CAN DO ANY COGNITIVE TASK HUMANS CAN DO

- AI ARCHIEVE RATIONALITY WHEN:
  - PERCIEVING THE ENVIORMENT
  - REASONING ON WHAT IS PERCEIVED
  - DECIDING WHAT THE BEST ACTION IS
  - ACTING ACCORDINGLY
  - (MODIFY THE ENVIORMENT)
- AI ARCHIEVE INTELLIGENCE WHEN:
  - ITS ACTION ARE APPROPRIATE FOR ITS GOALS
  - FLEXIBLE TO CHANGING ENVIORMENTS/GOAL
  - LEARN FROM EXPERIENCE
  - MAKE APPROPRIATE CHOICES GIVEN ITS PERCEPTUAL AND COMPUTATIONAL LIMITATION

TECHNICAL ARTIFACT

- OBJECT RELATED TO THE WORLD OF PHYSICAL OBJECTS - SOCIAL OBJECTS PART OF A SYSTEM
- NOT CHARACTERIZED FULLY IN TERMS OF INTRINSIC PHYSICAL PROPERTIES ALONE BUT ITS RELATIONAL/EXTRINSIC PROPERTIES

PHYSICAL OBJECTS DESIGNED BY HUMANS THAT HAVE A FUNCTION AND A USE PLANE

HYBRID OBJECTS WITH
- PHYSICAL FEATURES: ARE INTRINSIC FEATURES
- FUNCTIONAL FEATURES: HUMAN INTENTIONS OR PRACTICES OF INTENTIONAL HUMAN ACTIONS

AI AS SOCIO-TECHNICAL SYSTEM (KROES, 2006)

SYSTEM THAT DEPENDS ON TECHNICAL HARDWARE, HUMAN BEHAVIOUR AND SOCIAL ISTITUTIONS FOR THEIR PROPER FUNCTIONING

NON-TRADITIONAL SOCIOTECHNICAL SYSTEM FOR AI

(AI is not a traditional sociotechnical system)

- TECHNICAL ARTIFACTS
- ARTIFICIAL AGENTS
- TECHNICAL NORMS

AI PROPERTIES AS NON-TRADITIONAL SOCIOTECHNICAL SYSTEM (FLORIDI-SANDERS 2004)

- INTERACTIVITY: THE AGENT AND ITS ENVIRONMENT (CAN) ACT UPON EACHOTHER
- AUTONOMY: AGENT IS ABLE TO CHANGE STATE WITHOUT DIRECT RESPONSE TO INTERACTION. CAN PERFORM INTERNAL TRANSITION TO CHANGE ITS STATE (NON-DETERMINISTIC)
- ADAPTIVITY: AGENT'S INTERACTION CAN CHANGE THE TRANSITION RULES BY WHICH IT CHANGES STATE. LEARN FROM EXPERIENCE

4 TYPES OF AUTONOMY

- PERSONAL AUTONOMY: CAPACITY TO FORM PERSONAL VALUES, GOALS AND ULTIMATE ENDS
- MORAL AUTONOMY: REFLECTING OWNS MORAL PRINCIPLES OR ETCHICAL CONVINCTIONS
- RATIONAL AUTONOMY: ACTION ON THE BASIS OF THE WEIGHTIEST REASONS
- AGENTIAL AUTONOMY: IDENTIFY CERTAIN BEHAVIOUR AS GENUINE ACTION (NOT DETERMINED BY EXTERNAL FACTORS)

(AI percieves the environment based on situations ⇒ decide what is the best action)

HUMAN-TECHNOLOGIES INTERACTION

(most widely accepted view of technology: technology = tool)

1) INSTRUMENTALIST THEORY: TECHNOLOGY IS JUST A TOOL STANDING READY TO SERVE THE AIMS OF USERS. IT IS NEUTRAL. TECHNOLOGICAL ARTIFACTS ARE EVALUATED ON THE BASIS OF THE EMPLOYMENTS DECIDED BY ITS HUMAN DESIGNER OR USER.

(human have goal technology is a mean to reach it)

- THE USE OF TECHNOLOGY (NOT THE TECHNOLOGY ITSELF) IS MORALLY GOOD OR BAD (guns don't kill people, people kill people)
- TECHNOLOGICAL ARTIFACTS DO NOT CONTAINS VALUES. PEOPLE HAVE VALUES

GUNS DON'T KILL PEOPLE, PEOPLE KILL PEOPLE
- GUNS DESIGNED FOR THE PURPOSE OF RELEASING DEADLY PROJECTILES OUTWARDS
- DIFFICULT TO FIND ANOTHER UTILITY (EXCEPT EXCEPTIONAL USE) - MADE TO HURT PEOPLE
- IF YOU HAVE A GUN THE WORLD CHANGE SHAPE, ANIMALS = OBJECTS = POTENTIAL TARGETS

2) MEDIATION THEORY (we shape technology and technology shape us)
- TECHNOLOGIES NOT ONLY INSTRUMENTS BUT HAVE MEDIATING ROLE. SHAPE OUR RELATIONSHIP W/ THE WORLD.
- IS NOT NEUTRAL

4 TYPES OF INTERACTIONS

1) EMBODIMENT RELATIONS: USE TECHNOLOGIES BUT WE ARE NOT AWARE OF IT (DRIVING THE CAR, HUMANS SENSES EXTENDED TO THE CAR (OR GLASSES))
2) HERMENEUTICS RELATIONS: TECHNOLOGY IS VISIBLE AND THE WORLD IS INTERPRETED BY THE USER THROUGH TECHNOLOGY (A CLOCK REPRESENT TIME, BUT IS ALSO A PARTICULAR INTERPRETATION OF TIME)
3) ALTERITY RELATIONS: WE ARE INTERACTING WITH SOMETHING OTHER. QUASI-OTHER. (A HUMANOID ROBOT. WE SEE IT IN A SIMILAR WAY AS A HUMAN/PET. IS NEITHER A MERE TOOL NOR AN OBJECT PART OF THIS WORLD) (someone/something with which I interact)
4) BACKGROUND RELATIONS: TECHNOLOGY STRUCTURES AND MEDIATES OUR EXPERIENCE BUT IT REMAINS IN THE BACKGROUND (THERMOSTAT)

OVERPASSES IN LONG ISLAND

- BUILD OVERPASS LOW ⇒ LIMIT ACCESS TO AFRICAN-AMERICAN = TECHNOLOGY EMBODY SOCIAL RELATIONS
- SOCIETIES CHOOSE STRUCTURES FOR TECHNOLOGY THAT INFLUENCE HOW PEOPLE ARE GOING TO WORK, COMMUNICATE... FOR A LONG TIME

TECHNOLOGICAL INNOVATIONS ARE SIMILAR TO LEGISLATIVE ACTS OR POLITICAL FOUNDINGS THAT WILL WILL ENDURE OVER MANY YEARS → NEED TO GIVE ATTENTION TO BUILDING STREETS, CREATING TELEVISION...

THE ISSUE THAT UNITE/DIVIDE A SOCIETY ARE NOT ONLY IN ISTITUTIONS BUT ALSO IN TANGIBLE ARRANGEMENTS SUCH AS STEEL AND CONCRETE

(AA = ARTIFICIAL AGENT)

BIASES

EXAMPLE OF JOB RECRUITMENT. AI TRAINED OF ACTUAL HISTORICAL DATA. WHITE MALE MOST IMPORTANT PREDICTOR FOR TAKING THE JOB

→ HUMAN BIASES WERE TRANSFERRED AND AMPLIFIED BY THE AI

BIASES CAN ARISE

- ALL STAGES OF DESIGN, TESTING AND APPLICATION
- (TO FOCUS ON DESIGN)
- SELECTION OF THE TRAINING SET
- IN THE TRAINING SET ITSELF (UNREPRESENTATIVE/INCOMPLETE)
- IN THE ALGORITHM
- IN THE DATASET THE ALGORITHM IS GIVEN ONCE IS TRAINED
- IN DECISION BASED ON FAKE CORRELATION
- IN THE GROUP THAT CREATES THE ALGORITHM

THE MIRROR VIEW - COECKELBERG (2020)

- USE A DATASET THAT MIRRORS THE REAL WORD. ALGORITHM MAY MODEL EXISTING BIAS, NOT A PROBLEM FOR DEVELOPERS

VS

- BIAS AND DISCRIMINATION ARE UNFARE, ONE SHOULD CHANGE THE DATASET TO PROMOTE AFFIRMATIVE ACTION

VALUE SENSITIVE DESIGN

IF X IS VALUABLE ⇒ THEN ONE HAVE REASONS FOR A POSITIVE RESPONSE (PRO ATTITUDE OR PRO BEHAVIOUR)

- VALUE SENSITIVE DESIGN: WE CAN SOMEHOW DESIGN MORAL VALUES IN TECHNICAL ARTIFACTS SO THAT THEY CAN EMBODY VALUES ⇒ TECHNICAL ARTIFACTS MORALLY VALUE-LADEN BY DESIGNING THEM IN A WAY
- DEONTIC NOTIONS: DETERMINE THE RIGHTNESS/WRONGNESS OF ACTIONS. ENTITY EMBODIES A VALUE IF THERE ARE REASONS FOR PRO-ATTITUDE/PRO-BEHAVIOUR TOWARD THAT ENTITY

→ VALUES EVALUATES TECHNOLOGICAL ARTIFACTS IN TERMS OF GOODNESS AND BADNESS. 3 TYPES OF VALUES

1) INTENDED VALUES: THE VALUES WHICH DESIGNERS AIM TO EMBODY IN THEIR DESIGN AND WHICH THEY HOPE TO BE REALIZED IN PRACTICE
2) REALIZED VALUE: VALUE THAT IS REALIZED BY A TECHNICAL ARTIFACT IN A PRACTICAL CONTEXT
3) EMBODIED VALUE: POTENTIAL TO REALIZE A VALUE IN A APPROPRIATE CONTEXT

- TECHNICAL NORM: COMPUTE CODES THAT REGULATE THE BEHAVIOUR OF AN INTERACTIONS OF AA, TWO WAYS
- OFFLINE DESIGN: NORM ARE SPECIFIED AND ENCODED IN THE AGENT BY HUMANS
- AUTONOMUSLY DISCOVER, INVENT OR SPREAD NORMS: NORMS CAN BE PICKED FROM THE ENVIORMENT IN DIFFERENT WAYS OR FROM INTERACTIONS WITH AGENTS AND/OR HUMANS

→ VAN DE POEL: TECHNICAL NORMS (N) EMBODY VALUE (V) IF.

- N EMBODIES V
- THE EXECUTION OF N WITHIN THE SYSTEM IS CONDUCTIVE TO V

MORAL STATUS OF AI

SIMILAR TO TECHNICAL ARTIFACTS, BUT HAVE AUTONOMY, ADAPTIVITY AND INTERACTIVITY

ARTIFICIAL AGENTS ≠ HUMANS AGENT

• CAN EMBODY VALUES	• CAN NOT EMBODY VALUES
• CAN NOT (AA DO NOT HAVE INTENTIONALITY)	• CAN EMBED VALUES IN OTHER ENTITIES
• CAN ADAPT BEHAVIOUR FROM EXTERNAL INPUTS AND INTERACTIONS	• CAN ADAPT BEHAVIOUR FROM EXTERNAL INPUTS AND INTERACTIONS

WHAT IS AI CAPABLE OF DOING MORALLY SPEAKING?

- AGENT: BEING THAT CAN ACT
- AGENCY: EXERCISE/MANIFESTATION OF THIS CAPACITY
- MORAL AGENCY: INDIVIDUAL'S ABILITY TO MAKE MORAL CHOICES (BASED ON RIGTH-WRONG) RESPONSIBLE FOR THEM

HOW SHOULD WE TREAT AN AI

PATIENCY: QUALITY OR STATE OF BEING PATIENT OR PASSIVE

CATEGORIES OF ETHICAL AGENTS (MOOR)

PROBLEM: CREATING MORAL MACHINES IN THE SENSE THAT MACHINES THEMSELVES CAN MAKE ETHICAL PREDICTION

- ETHICAL IMPACTS AGENTS: ANY MACHINE THAT CAN BE EVALUATED FOR ITS ETHICAL CONSEQUENCES (PROBLEM: ALL ROBOTS HAVE ETHICAL IMPACTS)
- IMPLICIT ETHICAL AGENTS: MACHINES DESIGNED SO THAT THEY DO NOT HAVE NEGATIVE ETHICAL EFFECTS (PROBLEM: ALL ROBOTS SHOULD BE DESIGNED TO BE IMPLICIT ETHICAL AGENTS)
- EXPLICIT ETHICAL AGENTS: USING ETHICAL CATEGORIES AS PART OF THEIR INTERNAL PROGRAMMING
- FULL ETHICAL AGENTS: CAN MAKE EXPLICIT MORAL JUDGEMENTS AND ARE GENERALLY QUITE COMPETENT IN JUSTIFY SUCH DECISION

CHARACTERISTIC OF INTELLIGENT AGENT

- AUTONOMY: ENABLES TO PERFORM CERTAIN TASK ON THEIR OWN
- LEARNING ABILITY: ENABLES TO LEARN
- INTERACT WITH A NETWORK OF OTHER AGENTS: HUMAN/NON HUMAN (NATURAL/ARTIFICIAL)
- CAN LEARN NEW RULES INCREMENTALY
- GOAL-ORIENTED HABITS

MACHINES CAN NOT BE MORAL AGENTS (JOHNSON)

DON'T HAVE: MENTAL STATES, EMOTIONS AND FREE WILL

COMPUTER SYSTEMS HAVE MEANING, SIGNIFICANCE ONLY IN RELATION TO HUMANS. SOCIO-TECHNICAL SYSTEM

- NOT HAVE REQUIRED CAPACITY → NO MENTAL STATE, NO EMOTIONS, NO FREE WILL
- ARE PRODUCED BY HUMANS, ONLY HUMANS ARE ABLE TO DECIDE MORALLY
- AA DON'T HAVE INTENTION TO ACT BUT THEY ARE NOT NEUTRAL, THEY ARE MORAL ENTITIES THEY ARE PART OF THE MORAL WORLD NOT JUST BECAUSE OF THEIR EFFECT BUT BECAUSE WHAT THEY ARE AND DO

NATURAL PHENOMENA/ENTITY

- APPEAR IN NATURE INDEPENDENT OF HUMAN BEHAVIOUR
- HAVE A FUNCTIONALITY (BUT NOT DESIGNED BY HUMANS)
- DO NOT HAVE INTENTIONALITY

TECHNOLOGY

- A TOOL FOR HUMAN ENDS, DESIGNED BY HUMANS TO BEHAVE IN A SPECIFIC WAY

HUMAN-MADE OBJECTS

- HAVE A FUNCTIONALITY
- HAVE BEEN INTENTIONALITY DESIGN

IF NO DISTINCTION

- IMPOSSIBLE TO DISTINGUISH THE EFFECTS OF HUMAN BEHAVIOUR ON THE WORLD
- IMPOSSIBLE TO KNOW THE HUMAN CONTRIBUTION TO THE WORLD
- IMPOSSIBLE FOR HUMANS TO COMPREHEND THE IMPLICATION OF THEIR NORMATIVE CHOICES ABOUT THE FUTURE

DISTINCTION BETWEEN ARTIFACTS AND TECHNOLOGY

- A PIECE OF JEWELRY: CREATED BY HUMANS FOR THE SPECIFIC PURPOSE TO BE WORN FOR DECORATIVE REASONS. IT DOES NOT UTILIZE SCIENTIFIC KNOWLEDGE OR TECHNIQUES TO PERFORM TASKS OR SOLVE PRACTICAL PROBLEMS
- A SCULPTURE (PURPOSE OF ARTISTIC EXPRESSION)

- ARTIFACTS: PRODUCT OF HUMAN CONTRIVANCE, DOES NOT EXISTS WITHOUT SYSTEMS OF KNOWLEDGE
- HAVE MEANING ONLY IN THE CONTEXT OF HUMAN SOCIAL ACTIVITY
- ARE ABSTRACTION FROM REALITY/THE SYSTEM
- TO DEFINE AN ARTIFACT WE MUST DO A MENTAL ACT OF SEPARATING THE OBJECT FROM THE CONTEXT

- SOFTWARE PROGRAM: NOT PHYSICAL OBJECT (SO NO ARTIFACT) BUT THEY USE SCIENTIFIC KNOWLEDGE TO PERFORM TASKS OR SOLVE PROBLEMS
- MANUFACTURING PROCESS (THE PROCESS OF PRODUCING IS NOT A PHYSICAL OBJECT)

- TECHNOLOGY: SOCIO-TECHNICAL SYSTEM
- APPLICATION OF SCIENTIFIC KNOWLEDGE, TOOLS AND TECHNIQUES TO PRACTICAL TASKS AND PROBLEMS
- CAN INCLUDE BOTH PHYSICAL OBJECTS (MACHINES, DEVICES, ...)
- CAN INCLUDE INTANGIBLE PROCESS (ALGORITHMS, SOFTWARE, ...)

GIVE AI PRINCIPLES, CAN BE BETTER THAN HUMANS AS MORAL REASONING (NOT CARRIED AWAY BY EMOTIONS). AGAINST THE IDEA

- MORAL RULES OFTEN CONFLICTS
- MORALITY CAN NOT BE REDUCED TO FOLLOWING RULES, NOT JUST A MATTER OF RATIONALITY

STRUCTURE OF INT. AGENT:

- ARCHITECTURE: MACHINERY/DEVICE OF ACTUATORS - SENSORS
- AGENT FUNCTION: ACTIONS ARE MAPPED FROM A CERTAIN SEQUENCE
- AGENT PROGRAM: PRODUCTION OF THE AGENT FUNCTION

CONDITIONS FOR MORAL AGENCY

1) • AGENT WITH INTERNAL STATE: CONSISTS OF DESIRES, BELIEFS AND OTHER INTENTIONAL STATES. CONSITUTE A REASON TO ACTING
   • ONE INTENTIONAL STATE IS A MENTAL STATE (INTENTIONAL STATE + INTENT/REASON TO ACT)
2) • THERE IS AN OUTWARD EMBODIED EVENT. THE AGENT DO SOMETHING AND MOVE HIS BODY IN SOME WAY
3) • THE INTERNAL STATE IS THE CAUSE OF THE OUTWARD EVENT (THE MOVEMENT OF THE BODY IS RATIONALLY DIRECTED AT SOME STATE OF THE WORLD)
4) • THE OUTWARD BEHAVIOUR HAVE AN OUTWARD EFFECT
5) • THE EFFECT HAVE TO BE ON A RECIPIENT THAT CAN BE HARMED OR HELPED

- A COMPUTER SATISFY 2-5 BUT MACHINE DO NOT HAVE THE FREEDOM OR HUMAN INTENTIONALITY THAT WOULD MAKE THEM MORALLY RESPONSABLE
- COMPUTER SYSTEM DO NOT HAVE INTENDING TO ACT, DO NOT HAVE HUMAN INTENTIONALITY AND THIS IS THE KEY TO UNDERSTANDING THE MORAL CHARACTER OF COMPUTER SYSTEMS → CAN NOT BE MORAL AGENTS

→ SYSTEMS HAVE INTENTIONALITY. CONNECTED TO INTENTIONALITY OF THE USER AND OF THE DESIGNER. SYSTEM INTENTIONALITY IS LATENT WITHOUT USER INTENTIONALITY

- DEPENDENCE OF COMPUTER SYSTEM BEHAVIOUR ON HUMAN BEHAVIOUR
- SINCE COMPUTER SYSTEMS HAVE BUILD-IN INTENTIONALITY ONCE DEPLOYED THEY CAN BEHAVE INDEPENDENTLY AND WITHOUT HUMAN INTERVENTION

- COMPUTER SYSTEM ARE NOT-NEUTRAL AND CLOSE (BUT NOT) MORAL AGENTS. THEY ARE MORAL ENTITIES (SO THEY ARE PART OF THE MORAL WORLD)
- COMPUTER ARE NOT MORAL AGENTS: NO MENTAL STATE - INTENDING TO ACT. BUT THEY ARE CLOSE TO BEING MORAL AGENTS: INTENTIONALITY CREATED - USED FORMS OF INTENTIONALITY/EFFICACY. THEY ARE PART OF THE MORAL WORLD: BECAUSE OF THEIR EFFECTS - WHAT THEY ARE AND WHAT THEY DO

INTENTIONALITY OF COMPUTER SYSTEM

- INTENTIONALITY OF THE COMPUTER SYSTEM DESIGNER
- INTENTIONALITY OF THE SYSTEM
- INTENTIONALITY USER

ALL HAVE INTENTIONALITY-EFFICACY. THEY BEHAVE FROM NECESSITY WHEN CREATED. COMPUTER SYSTEM COMPONENTS IN MORAL ACTION

FUNCTIONAL MORALITY (WALLACH-ALLEN)

HOW CAN WE CREATE AN ARTIFICIAL MORAL AGENT?

- MACHINES WITH AUTONOMY AND SENSITIVITY TO VALUES HAVE FUNCTIONAL MORALITY

FUNCTIONAL MORALITY (FM)

- AUTOPILOTS AND ETHICAL DECISIONS SUPPORT SYSTEMS
- IS TECHNOLOGICALLY POSSIBLE
- CAPACITY OF MACHINES FOR ASSESSING AND RESPONDING TO MORAL CHALLENGES
- MACHINE MUST HAVE AUTONOMY AND SENSITIVITY TO VALUES

OPERATIONAL MORALITY (OM)

- GUNS WITH CHILDPROOF MECHANISM, IT LACKS OF AUTONOMY AND SENSITIVITY BUT ITS DESIGN EMBODY VALUES OF NRA CODE OF ETHICS TO ENDORSE

DIFFERENT ETHICAL THEORIES IN THE PROCESS OF DESIGN

DIFFERENT ETHICAL THEORIES AND APPROACHES TO PRODUCE MACHINE CAPABLE OF DECISION-MAKING

- **TOP-DOWN APPROACH:**
  - SET OF RULES THAT CAN BE TURNED IN A ALGORITHM
  - **UTILITARIANISM:** MAXIMISE THE TOTAL AMOUT OF UTILITY IN THE WORLD. BEST ACTION THE ONE THAT MAXIMISE AGGREGATE UTILITY (MEASURE OF HAPPINESS/WELL BEING)
    - HEAVY COMPUTATION: NEEDS TO WORK OUT MANY/ALL THE CONSEQUENCES IN ORDER TO RANK ACTIONS MORALLY
    - HOW TO CONSTRUCT A FUNCTION THAT CAN EVALUATE WITH THE CORRECT WEIGTH THE PRESENT BENEFITS AGAINST THE FUTURE ARMS?
    - HOW TO ASSIGN NUMBERS TO SOMETHING SUBJECTIVE LIKE HAPPINESS, PLEASURE OR DESIRABILITY?
  - **DEONTOLOGICAL:** (WHAT SHOULD WE DO? OUR DUTY) EVALUATES THE ETHICAL CORRECTNESS OF ACTIONS ON THE BASIS OF CHARACTERISTIC THAT AFFECT THE ACTION IN ITSELF
    - ANY LIST OF DUTIES MIGTH HAVE INTERNAL CONFLICT
    - DESIGNER NEED TO BE SURE THAT THE RULES ARE ACTIVATED WHEN A SITUATION REQUIRES THEIR APPLICATION
    - FORMULATE AN ARCHITECTURE FOR MANAGING SITUATIONS WHEN RULES CONFLICTS
- **BOTTOM-UP APPROACH:** (TURING CHILD MIND: SUBJECT TO THE COURSE OF EDUCATION ⇒ OBTAIN THE ADULT BRAIN)
  - MIMIC CHILD'S MORAL DEVELOPMENT
  - DEVELOPMENT NOT EXPLICITLY GUIDED BY ANY ETHICAL THEORY
  - DESIGN METHODS FOR CARRYING OUT DISCRETE TASKS ⇒ MORE COMPLEX ACTIVITIES/GREATER AUTONOMY
  - NEED SOME METRIC FOR EVALUATION IF A DEVICE IS CAPABLE OF MAKING THE APPROPRIATE MORAL DECISION IN A PARTICULAR SITUATION
  - ANTHROPOCENTRISM
  - SLAVE ETHICS

GUNKEL VS FUNCTIONALIST APPROACH

- **MERGIN TOP-DOWN AND BOTTOM-UP APPROACH:**
  - **MORAL APPROACH:** COECKELBERG PROPOSED QUASI-MORAL ROBOTS: APPEAR TO BE MORAL
    - NO PROOF THAT ANOTHER PERSON HAVE MENTAL STATES/CONSCIOUSNESS. BUT WE INTERPRET THE OTHER'S APPERANCE/BEHAVIOUR = EMOTION. THE OTHER HAVE VIRTUAL SUBJECTIVITY/QUASI-SUBJECTIVITY, INTERACTING AS IF OUR APPERANCE/BEHAVIOUR IN THEIR CONSCIOUSNESS
    - SAME WITH ROBOTS: IP SUFFICIENTLY ADVANCED (INITATE SUBJECTIVITY/CONSCIOUSNESS IN A CONVINCING WAY) THEY BECOME/APPEAR THE QUASI-OTHER
  - **RELATIONAL APPROACH:** COECKELBER. MORAL STANDING: NOT AN OBJECTIVE PROPERTY. HUMANS ATTRIBUTE THE MORAL STATUS ⇒ STATUS OF THE OBJECT OF MORAL STANDING. DEPENDENT FROM HUMAN SUBJECTIVITY ⇒ ROBOT CAN APPEAR IN DIFFERENT WAYS TO DIFFERENT PEOPLE IN DIFFERENT SITUATIONS/CONTEXT

WHAT MATTERS MORALLY SPEAKING, IS HOW THE ENTITY APPEARS

ETHICAL ATTENTION SHIFTED

- FROM ONTOLOGY (PROPERTIES OF AN ENTITY) TO EPISTEMEOLOGY (OUR KNOWLEDGE OF THE ENTITY)
- FROM OBJECT TO SUBJECT
- FROM WHAT THINGS REALLY ARE TO HOW WE LOOK AT THINGS

MINDLESS MORALITY (FLORIDI-SANDERS)

MAKE MORAL AGENTS SUFFICENT LEVEL OF:

- **INTERACTIVITY** → RESPONSE TO THE STIMULUS BY CHANGING STATE
- **AUTONOMY** → ABILITY TO CHANGE STATE W/OUT STIMULUS
- **ADAPTIVITY** → ABILITY TO CHANGE THE "TRANSITION RULE" BY WHICH STATE IS CHANGED
- CAPABLE OF MORAL QUALIFIABLE ACTIONS

AN AGENT IS MORALLY GOOD IF ALL HIS ACTIONS RESPECT THIS THRESHOLD (AND MORALLY EVIL IF SOME ACTIONS VIOLATES IT)

WHAT IS AUTONOMY?

- FREEDOM FROM EXTERNAL CONTROL/INFLUENCE = INDEPENDENCE
- CAPACITY TO MAKE AN INFORMED/UNCOERCED DECISION = FREE DECISION
- KANT & RATIONAL AUTONOMY. ONE POSSESS THE MOTIVATION TO GOVERN THEIR OWN LIFE

SULLINS REQUIREMENTS FOR FULL MORAL AGENCY

ROBOTS ARE MORAL AGENTS WHEN THERE IS A REASONABLE LEVEL OF ABSTRACTION UNDER WHICH WE MUST GRANT THAT THE MACHINE HAS AUTONOMOUS INTENTIONS AND RESPONSIBILITIES

- **AUTONOMY:** (AUTONOMOUS FROM PROGRAMMER/OPERATOR/USER) MACHINE NOT UNDER THE DIRECT CONTROL OF ANY OTHER PROGRAMMER OR USER. WHEN THE AGENCY CAUSES HARM OR GOOD IN A MORAL SENSE WE CAN SAY THE MACHINE HAVE MORAL AGENCY
  IF THE ROBOT ACTIONS IS EFFECTIVE IN ACHIEVING THE GOALS/TASKS OF THE ROBOT ⇒ THEN THE ROBOT HAS EFFECTIVE AUTONOMY
- **INTENTIONALITY:** (DO NOT PROVE IT HAS INTENTIONALITY) IF THE COMPLEX INTERACTION OF THE ROBOT'S PROGRAMMING AND ENVIORNMENT CAUSES THE MACHINE TO ACT IN A MORALLY HARMFUL OR BENEFICIAL WAY AND THE ACTIONS SEEMS DELIBERATE AND CALCULATED THEN THE MACHINE IS A MORAL AGENT
- **RESPONSIBILITY:** ROBOT BEHAVES RESPONSIBLE TO OTHER MORAL AGENTS. IF A ROBOT BEHAVES IN THIS WAY. FULLFILLS SOCIAL ROLE OF SOME RESPONSABILITIES ⇒ IT HAS THE DUTY TO CARE FOR ITS PATIENTS ⇒ MORAL AGENT