TURING TEST (IMITATION GAME)

↗ COMPUTING MACHINERY AND INTELLIGENCE

- 1950, TURING. 3 PLAYERS (1 MACHINE(M) AND 2 HUMANS(H))
- H1 HAVE TO GUESS WHO IS THE OTHER HUMAN BY ASKING QUESTIONS TO M AND H_2
- H_2 HAVE TO HELP H_1, M HAVE TO FOOL H_1 TO MAKE H_1 BELIVE THAT M IS H_2 INSTEAD
- REPLACES A PSYCHOLOGICAL OR COGNITIVE UNDERSTANDING (CAN A MACHINE THINK?) WITH AN EMPIRICAL OR BEHAVIOURISTIC UNDERSTANDING (IT FOCUS ON THE INTERNAL BEHAVIOUR OF A MACHINE RATHER THAN ITS INTERNAL COGNITIVE PROCESSES)

TURING TEST DOES NOT SEEK TO UNDERSTAND HOW A MACHINE THINKS OR WHAT PROCESSES ARE TAKING PLACE INSIDE OF IT. IT EVALUATES A MACHINE ABILITY TO FOOL A HUMAN, BUT IT DOES NOT MEANS THAT IT BEHAVES OR THINK LIKE A HUMAN

- LIMITATIONS:
  - FOOLING DOES NOT MEANS THAT A MACHINE IS ACTUALLY THINKING OR EXHIBITING INTELLIGENT BEHAVIOUR IN THE WAY THAT HUMAN DOES
  - CHAUVINISTIC: RECOGNISE INTELLIGENCE ONLY IN THINGS THAT ARE ABLE TO HAVE A CONVERSATIONS WITH HUMANS
  - NOT SUFFICIENT DEMANDING: A MACHINE CAN PASS THE TURING TEST FOR REASONS OTHER THAN THE POSSESSION OF INTELLIGENT
- SUBJECTIVE TEST: RELIES ON THE JUDMENT OF THE EVAWATOR (H_4)
- LIMITING TO DIGITAL COMPUTERS: IF IT WAS A GOOD TEST IT SHOULD BE APPLICABLE TO OTHERS (ANIMAL, ALIENS, ANALOG COMPUTERS)

- WEAK AI: MIGTH IMITATE HUMANS AND PASS THE TURING TEST
- STRONG AI: HAVE CONSCIOUSNESS, INTENTIONALITY AND UNDERSTANDING

↗ FOR SEARLE

CHINESE ROOM (SEARLE)

- THOUGHT EXPERIMENT
- ROOM WITH A SET OF ISTRUCTION IN ENGLISH ABOUT HOW TO RESPONDS IN CHINESE (THAT YOU DON'T KNOW)
- RECIEVE A CHINESE NOTE, USE ISTRUCTION TO RESPONDS, PEOPLE OUTSIDE THE ROOM WRONGLY BELIVE THAT HE SPEAKS CHINESE
- AIM TO SHOW THAT THE TURING TEST IS INADEQUATE BECAUSE IMITATIONS IS NOT SUFFICIENT FOR UNDERSTANDING AND UNDERSTANDING IS NOT PRESENT IN WEAK AI

ETHICS:

- IS ABOUT WHAT ONE SHOULD DO, ALL THINGS CONSIDERED
- CONCERNED WITH WHAT PEOPLE SHOULD DO OR SHOULDN'T DO. IS ABOUT DUTIES AND PERMISSION

- DESCRIPTIVE ETHICS: STUDY OF HOW PEOPLE ACTUALLY BEHAVE AND MAKE MORAL DECISION, DESCRIBE AND EXPLAIN MORAL BELIEFS AND PRACTICES AND HOW THEY VARY ACROSS DIFFERENT CULTURES AND SOCIETY. DOES NOT EVALUATE RIGHTNESS OR WRONGNESS OF MORAL BELIEFS AND PRACTICES, INSTEAD IT DESCRIBE HOW THEY ARE PRACTICES

DESCRIBE AND EXPLAIN MORAL BELIEFS AND PRACTICES

- NORMATIVE ETHICS: STUDY OF MORAL PRINCIPLES AND VALUES AND HOW THEY SHOULD BE APPLIED TO MORAL PROBLEMS AND DECISIONS. DESCRIBE AND EVALUATES WHAT ACTIONS AND BEHAVIOURS ARE RIGTH OR WRONG, AND DEVELOPING MORAL THEORIES AND PRINCIPLES THAT CAN GUIDE MORAL JUDGMENTS AND DECISION-MAKING

↗ AKA FIRST-ORDER MORAL THEORY. GIVES A CONCRETE RECIPE ON WHAT TO DO ON CONCRETE ISSUE

EVALUATING AND PRESCRIBING WHAT ACTIONS AND BEHAVIOURS ARE RIGHT OR WRONG

- META ETHICS: STUDY OF QUESTIONS ABOUT ETHICS
  - SEMANTIC: WHAT IS THE MEANING OF GOOD, RIGTH, CORRECT?
  - EPISTEMOLOGICAL: HOW?
  - ONTOLOGICAL: DO SUCH VALUES EXIST?

↗ AKA SECOND ORDER MORAL THEORY

- APPLIED ETHICS: APPLICATION OF CERTAIN FIRST ORDER MORAL TO SPECIFIC CONCRETE ISSUES

↳ AKA THIRD ORDER MORAL THEORY

- EMOTIVISM (STEVENSON): WHEN I SAY "ABORTION IS WRONG" I CONVEY AN EMOTION "BOO ABORTION" ⇒ DON'T LIKE IT, OR "W ABORTION"
- NON-COGNITIVIST POSITION (HARE): REFINE EMOTIVISM, DISTINGUISHED BETWEEN THE CONTEXT OF A WORD AND ITS MOOD (EITHER DESCRIPTIVE OR EXPRESSIVE). THE POSITION ACCORDING TO WHICH MORAL SENTENCES ARE NOT TRUTH FUNCTIONAL AND DO NOT EXPRESS DESCRIPTIVE KNOWLEDGE AS IT WERE
- EXPRESSIVISM: MORAL STATEMENTS DO NOT DESCRIBE OBJECTIVE TRUTHS BUT RATHER EXPRESS THE SPEAKER'S PERSONAL FEELINGS OR EVALUATIONS

FOUR FAMILY of FIRST ORDER MORAL THEORY

DIFFEREN THEORIES, TRY TO ANSWER TO

- WHAT SHOULD WE DO?
- WHAT IS THE RIGTH THING TO DO?
- WHAT IS GOOD?

GIVE DIFFERENT ANSWERS

DEONTOLOGY

↗ THERE CAN BE MORE THAN 1 DUTY

- DEONTOLOGY: WHAT SHOULD WE DO? OUR DUTY.
  - THERE IS A PRINCIPLE THAT YOU SHOULD FOLLOW THAT ENCODES OUR DUTY.
  - EACH DEONTOLOGICAL THEORY WILL TELL YOU WHAT THE DUTY IN QUESTION IS AND WHAT PRINCIPLE TO FOLLOW (10 COMMANDMENTS)
  - MORAL DILEMMAS: EXAMPLE OF HIDING SOMEONE AND DO NOT LIE FROM THE KILLER, WHAT TO DO?
  - HIERARCHY OF PRINCIPLES: ONE PRINCIPLE MORE IMPORTANT THEN THE OTHER
  - MONIST (THEORY OF VALUE) ONLY ONE PRINCIPLE IMMANUEL KANT + GOLDEN RULE
  - PLURALISM (WILLIAM DAVIS ROSS): THERE ARE MULTIPLE DUTIES (DUTY OF BENEFICENCE, NOT MALEFICENCE...) THERE IS A HIERARCHY BETWEEN THOSE PRINCIPLES BUT IN A PARTICULAR SITUATION YOU SHOULD FIGURE OUT WHAT YOUR DUTY IS

VIRTUE ETHICS

- YOU SHOULD DO WHAT'S VIRTUOUS
- LESS COMMON
- IF YOU ARE A SOLDIER YOU SHOULD BE A GOOD ONE (OBEY THE ORDERS...)

PARTICULARISM - CONTEXT TO CONTEXT

- MORAL PARTICULARISM: IS NOT THE CASE THAT THERE IS A GENERAL PRINCIPLE OF ACTION
- IN ORDER TO DECIDE WHAT TO DO YOU HAVE TO GO THROUGTH A BALANCING OR WEIGHING OF THOSE REASONS IN FAVOR AND REASONS AGAINST
- THE REASON CAN SWITCH POLARITY IN DIFFERENT SITUATIONS

CONSEQUENTIALISM

- WHAT YOU SHOULD DO IN A SITUATITION DOES NOT DEPENDS ON THE ACTIONS OF A PARTICULAR PRINCIPLE BUT RATHER ON THE CONSEQUENCES OF YOUR ACTIONS
- DIFFERENT THEORIES ON HOW TO EVALUATE THE CONSEQUENCES OF YOUR ACTIONS

1) HEDOISM: ORIGINAL THEORY of CONSEQUENTIALISM: ACTION ARE EVALUATED ON THE AMOUNT OF PLEASURE THEY CAUSE AND ON THE AMOUNT of PAIN THEY AVOID
2) UTILITARIANISM: YOU SHOULD TAKE THE BEST COURSE OF ACTION?
   - 2a) ACT UTILITARIANISM: YOUR SINGLE ACT SHOULD MAXIMISE UTILITY. ACTION DONE BY A SINGLE INDIVIDUAL TO REACH THE OVERALL HAPPINESS OF EVERYONE (IF JUST FOR ITSELF → EGOISM)
     - PROBLEM: IS REALLY HARD TO FORESEEN THE CONSEQUENCES OF YOUR ACTIONS IN A GLOBAL COMPLETE WAY
   - 2b) RULE UTILITARIANISM: YOU SHOULD DO THE ACTION THAT AGREE WITH THE PRINCIPLE THAT HAS THE BEST CONSEQUENCES. ACT DONE COLLECTIVELY BY FOLLOWING SOME RULES TO REACH AN OVERALL HAPPINESS FOR EVERYONE.
3) EGOISM: I SHOULD DO THE ACTIONS THAT HAVE THE BEST CONSEQUENCES FOR ME

ONTOLOGY - PART of METAETHICS: DOES MORAL FACTS AND VALUES EXISTS?

- THEORY REALISM (AKA OBJECTIVISM): MORAL FACTS AND VALUES EXISTS OUT THERE IN THE REAL WORLD AND THEY ARE OBJECTIVE.
  - MOORE OPEN QUESTION: AMONG OTHER THINGS PURPORTS TO ESTABLISH THAT GOOD(NESS) IS NOT REDUCIBLE TO ANY OTHER PROPERTY.
  - ANOTHER FORM OF REALISM CONSIDERS MORAL PROPERTY DISPOSITIONAL, SO NOT INDEPENDENT OF ANYTHING ELSE, BUT SOMEWHAT EMBEDDED IN THE FABRIC OF REALITY
- THEORY ANTI-REALISM (AKA RELATIVISM): MORAL FACTORS AND VALUES DO NOT EXISTS OUT THERE IN THE WORLD
  - THEY HAVE CERTAIN LEVEL OF SUBJECTIVE EXISTENCE, OR INTER-SUBJECTIVE EXISTENCE

EPISTEMOLOGY - PART of METAETHICS

- HOW WE ATTAIN (IF WE DO) MORAL KNOWLEDGE
- COGNITIVISTS: MORAL KNOWLEDGE IS EXACTLY AS ALL OTHER SORTS OF KNOWLEDGE (MOSTLY PROPOSITIONAL AND FACTIVE) ⇒ IT IS TRUTH-FUNCTIONAL
- NON-COGNITIVIST: MORAL KNOWLEDGE IS NOT PROPOSITIONAL BUT RATHER INVOLVING NON-COGNTIVE MENTAL STATES (EMOTION, DESIRES,...)
- HOW WE JUSTIFY MORAL BELIEFS? ⇒ INTUITIONISM

MORAL UNCERTAINTLY: WHAT SHOULD WE DO IF ONE IS NOT SURE WHICH FIRST ORDER MORAL THEORY IS CORRECT?

↳ DEONTOLOGY, VIRTUE ETHICS, PARTICULARISM, CONSEQUENTIALISM...

- ONE WEIGTHS THEIR SUBJECTIVE CREDENCE ON ANY SINGLE MORAL THEORY AND MULTIPLIES IT FOR THE VALUE OF THAT PARTICULAR OUTCOME ACCORDING TO THAT PARTICULAR MORAL THEORY

LIMITATIONS

- SUBJECTIVE: REFLECTS ONE'S CONVICTIONS AND BIASES RATHER THAN BEING OBJECTIVE. LIMITED TO THE INDIVIDUAL LEVEL
- WORK ONLY IN FAIRLY HOMOGENEOUS NORMATIVE CONTEXT. IS DIFFICULT TO APPLY TO A MORE HETEROGENEOUS LANDSCAPE, BECAUSE IT ASSUMES INTERTHEORETICAL COMPARABILITY OF THE VALUE/UTILITY FUNCTIONS, SOMETHING THAT CAN NOT BE ASSUMED FOR A DEONTOLOGICAL AND CONSEQUENTIALIST THEORY

ETHICS OF AI:

- IS **APPLIED ETHICS**: HOW CERTAIN FIRST-ORDER MORAL QUESTIONS APPLY TO A SPECIFIC DOMAIN (of AI) SOME PROBLEMS ON HOW TO PERFORM AI RESEARCH AND CREATE ALGHORITMS
- IS **DESCRIPTIVE ETHICS**: CERTAIN AI TECHNIQUES CAN BE USED TO GATHER AND ANALYZE ETHICAL DATA AND PREDICT PATTERNS
- IS **NORMATIVE ETHICS**: CAN INTELLIGENT BUT NON-HUMAN AGENTS BE MORAL PATIENTS? AND MORAL AGENTS?

- KNOW ETHICAL THEORIES AND THE EXTENT TO WHICH MACHINES CAN BE MORAL AGENTS/PATIENTS
- HOW CAN WE BUILD ETHICAL ARTIFICIAL AGENTS? COULD FOLLOW DIFFERENT MORAL PRINCIPLES

- **DEONTOLOGIST**: LAW OF ROBOTICS, THERE ARE CERTAIN PRINCIPLES THAT A MORAL AGENT MUST FOLLOW CATEGORICALLY
  - **LIMITATIONS**
    - MISINTERPRET THE RULES
    - LEAD TO UNFORESEEN SCENARIOS
    - THERE ARE ALREAD PROBLEMS WITH HUMAN ONTOLOGY. HARD TO GO FROM A GENERAL PRINCIPLE TO A CONCRETE IMPLEMENTATION
- **CONSEQUENTIALISM**: SOME ELEMENTS OF MACHINE LEARNING RESEMBLE ELEMENTS OF CONSEQUENTIALIST THEORY ALREADY
  - REWARD FUNCTION
  - EVALUATING POSSIBLE OUTCOMES
  - WORKING WITH PROBABILITY DISTRIBUTION
  - **LIMITATIONS**
    - TO WHAT EXTENT YOU ARE CLOSE ENOUGHT TO PREDICT THE CONSEQUENCES OF YOUR ACTIONS
    - HOW TO EVALUATE CONSEQUENCES
    - THE COST OF "RERUNNING THE PROGRAMME"?
    - + ALL THE PROBLEMS of CONSEQUENTIALISM
- **PARTICULARISM**: TRAIN A NETWORK USING ETHICAL DILEMMAS, AND THEN THE SYSTEM IS SUPPOSED TO SOLVE NEW ONES
  - DELHI: CHANGING THE GRAMMAR STRUCTURE CAN LEAD TO DRASTICALLY DIFFERENT OUCOMES
  - **LIMITATIONS**
    - ASSUMPTION IN THE TRAINING: ARE THERE CORRECT ANSWERS THAT A MACHINE COULD BE TRAINED ON? AND ACCORDING TO WHO AN ASWER IS CORRECT?
    - NEED TO SOLVE A LOT OF PROBLEMATIC CASES AND WE STILL WILL NOT BE SURE TO TRUST THE MACHINE
- **COMBINED APPROACHES: DEONTOLOGY + CONSEQUENTIALISM**
  - USE UTILITARIANISM UNTIL "SACRED VALUES" ARE CONCERNED
  - AT THAT POINT USE DEONTOLOGICAL TO BE LESS SENSITIVE TO THE UTILITY OF ACTIONS AND THE CONSEQUENCES
- **COMBINED APPROACH: THE "HYBRID APPROACH"**: COMBINE
  - TOP-DOWN: THEORY-DRIVEN REASONING
  - BOTTOM-UP: REASONING SHAPED BY EVOLUTION AND LEARNING

- SOCIETAL PROBLEM: NO AGREEMENT ON WHICH THEORY IS THE CORRECT ONE
  - DIFFERENT THEORIES ARE NOT ALWAYS COMPATIBLE
  - PROBLEM of MORAL UNCERTAINTY: WHAT SHOULD WE DO WHEN WE DO NOT KNOW WHAT THEORY IS CORRECT?
  - POSSIBLE SOLUTION: ETHICAL KNOB: IN A SELF-DRIVING CAR THE OWNER DECIDES WHAT MORAL THEORY TO IMPLEMENT, ONCE CHOOSEN THE CAR WILL ACT IN ACCORDANCE WITH IT TO ALL DILEMMA

EXISTENTIAL RISK AND LONGETERISM

- EXISTENTIAL RISK: EVENT THAT RESULTS IN INCREASING THE CHANGE OF HUMAN EXSTINTION OR UNCOVERABLE GLOBAL CATASTROPHE. NOT LIMITED TO AI (CLIMATE CHANGE, A NEW VIRUS, NUCLEAR DISASTER...)
- NEED TO MAKE EXPLICIT OUR VALUES

VS

- WHO CARES ABOUT LIFES IN 5000 YEARS? IS NOT WORTH TO CHANGE OUR LIFES FOR SCENARIOS FAR IN THE FUTURE
- LONGTERMINISM: FUTURE PEOPLE MATTERS AS MUCH AS THE PEOPLE ALIVE NOW
  - WE HAVE THE DUTY TO PRESERVE AND IMPROVE LIFE ON THIS PLANET BECAUSE THERE WILL BE MORE PEOPLE IN THE FUTURE THEN NOW
  - PRIORITY TO IMPROVE THE FUTURE RATHER THAN THE PRESENT

LIMITATIONS:

- LONG TERMINISM IS LINKED TO CONSEQUENTIALIST ETHICAL THEORIES HARDER TO ADOPT FOR PEOPLE ENDORSING OTHER
- HOW DO WE MEASURE FUTURE VALUE? LARGE THEORETICAL AND PRACTICAL DIFFICULTY
- A MORE DIRECT CRITICISM: DEPRIORITIZE OF CURRENT SERIOUS PROBLEMS

AUTONOMOUS WEAPONS SYSTEM

AUTONOMOUS WEAPONS SYSTEM: ALREADY EXISTS AND NOT NECESSARY LETHAL, AND NOT ILLEGAL ACCORDING TO THE RULES OF WAR

- LAW OF ARMED CONFLICT (LOAC) IS A COMPROMISE BETWEEN MORALITY AND THE REALITIES OF WAR
- LOAC IS NOT AS DEMANDING AS MORALITY, IS A SOURCE OF STANDARIZED SANCTIONING FOR BREACHES
- LOAC REQUIRES COMBATANTS AND COMMANDERS TO ABIDE BY ITS DICTATES + REQUIRES BREACHES TO BE PROSECUTED
- UNETICAL TO DEPLOY AUTONOMOUS SYSTEMS INVOLVING SOPHISTICATED ARTIFICIAL INTELLIGENCE IN WARFARE UNLESS SOMEONE CAN BE HELD RESPONSIBLE FOR THE DECISION THEY MAKE
- AWS: AVENGER WEAPONS SYSTEM, CAN SELECT AND ENGAGE A TARGET WITHOUT FURTHER INTERVENTION BY A HUMAN OPERATOR
  - ADVANCED MUNITIONS: ORDANCE THAT CAN SELECT AND ENGAGE TARGETS WITHOUT FURTHER INTERVENTION
  - COUNTER-ASSETS AWS: ENGAGE SOME FORM OF ENEMY ASSETS (RADAR INSTALLATION, ARMORED UNIT, AIRCRAFT, MISSILES...) MOSTLY ARE AUTONOMOUS MUNITIONS
  - POINT-DEFENSE AWS: WEAPONS PLATFORM DESIGNED TO PROTECT A SINGLE OBJECT OR VERY LIMITED AREAS (SHIP, TANK)
  - ANTI-PERSONELL AWS: TARGET INDIVIDUALS, SOME VARIANTS ARE DEPLOYED AS MINES

RESPONSABILITY GAP

- IF AWS ACT IN A UNFORESEEN OR UNPREDICTED WAY CAN DO THINGS THAT ARE A WAR CRIME. WHO IS RESPONSABLE?
- SOMETIMES EVEN TOO FAST TO BE STOPPED

- EACH AUTONOMOUS SYSTEM MUST BE EVALUATED ON ITS OWN TERMS
- TRULY UNPREDICTABLE SYSTEMS CREATE WORRIES BUT THESE WORRIES CAN BE MITIGATED

CURRENT USE of AI IN THE LAW

- LITIGATION DISCOVERY: AI TOOLS TO ACCESS EXISTING DATABASE AND ALLOW PRATICTIONERS TO REVIEW A LARGE NUMBER OF CONTRACTS
- DOCUMENT ASSEMBLY: PUT CONTRACTS AND OTHER LEGAL DOCUMENTS TOGETHER

USUALLY IS NOT PROBLEMATIC

- PREDICTION OF LEGAL OUTCOMES (COMPASS): USED BY JUDGES TO PREDICT WHETER OR NOT A PARTICULAR PERSON WAS LIKELY TO COMMIT A CRIME AGAIN (RECIDIVISM) BASED ON CERTAIN VARIABLES (SEX, ADRESS, INCOME...), THE DEFENDANTS COULD NOT ACCESS THE ALGORITHM THAT IS THE CORE OF THE PREDICTION SYSTEM
  - BIAS IN ALGORITHMIC DECISION MAKING, THE SYSTEM CREATES ITS PREDICTIVE MODEL BASED UPON PAST ARREST RECORDS PROVIDED BY THE POLICE
  - HUMAN BIASES ARE VERY LIKELY TO LEAD TO BIASES OUTPUT, EVEN IF THE ALGORITHM ITSELF IS UNBIASES
  - GARBAGE IN GARBAGE OUT
  - DEFERENCE TO AUTOMATED COMPUTERIZED DECISION-MAKING: GIVE THE WORK TO DO TO AN AI, RATHER THEN HUMANS BECAUSE THEY ARE NOT PERFECT AND CAN BE UNPRECISE AND MAKE MISTAKES

QUESTION OF TRANSPARENCY THAT SURROUNDS AI TECHNOLOGIES

- USE BY GOVERMENT OFFICIALS.
  - DISTRIBUTION OF GOVERNMENT BENEFITS: SOCIAL HOUSING AND UNEMPLOYED BENEFIT ALREADY USE AI
  - PREDICTIVE POLICING: ML TECHNOLOGIES USED TO DETECT PATTERNS FROM PAST CRIME DATA TO ATTEMPT TO PREDICT THE LOCATION AND TIME OF FUTURE CRIME ATTEMPTS

← MAY ALSO BE USED TO PROFILE PEOPLE WHO ARE MORE LIKELY TO COMMITT CERTAIN CRIMES

- AT THIS RATE IT CAN BE ABUSED
- UNCLEAR ABOUT SURVEILLANCE IN PUBLIC PLACES
- UNCLEAR FOR HOW LONG THE DATA WILL BE KEPT AND FOR WHAT USE

← FACIAL RECOGNITION TECHNOLOGY: EU AI ACT REGULATE THE USE OF FACIAL RECOGNITION SOFTWARE AND FORBID THE USE IN REAL TIME OR CAN BE USE ONLY IF THERE IS EVIDENCE OF A CRIME

↓

WOULD NOT PERMIT THE USE FOR PREVENTION OR PROFICING OF GROUPS OR INDIVIDUALS WHO ARE TAKEN TO BE CRIMINAL

- SMART CONTRACTS: USED FOR TRAINING
  - LEGAL-SELF HELP SYSTEM: PROVIDE ORDINARY USERS WITH ANSWERS TO BASIC LEGAL QUESTIONS (CHATBOT). MECHANICAL AI TO HELP YOU NAVIGATE IN THE KNOWLEDGE of LEGAL SYSTEM

← IF THERE IS A GLICH THE STOCK MARET LOSES BILIONS OF DOLLARS
- DUE TO THE OPACITY THE GLICH CAN NOT BE EXPLAINED

AI, CRIMINAL LAW AND THE LIABILITY GAP

CRIMINAL LAW IS ADRESSED ONLY TO HUMANS (AT LEAST SO FAR) BECAUSE HUMANS ARE THE ONLY ENTITY TO POSSES THE MENTAL STATE FOR AN ACT TO BE DEEMED CRIMINAL

- **CRIMINAL LAWS:** RULES ABOUT WHAT IS DEEMED MINIMALLY ACCEPTABLE BEHAVIOUR IN SOCIETY AND PUNISHMENT FOR BREAKING THOSE RULES
- (IN GENERAL) AN ACTION TO BE CONSIDERED A CRIME, THE ACTION ITSELF MUST BE CRIMINAL IN NATURE AND YOU HAVE TO PERFORM THE ACTION RECKLESSY OR WITH MALICE — ENSURE THE CONVICTED WAS DOING SOMETHING WRONG AND NOT JUST UNLUCKY IN SOME WAY
- **RESPONSIBILITY:** IS A PRECONDITION OF PUNISHMENT DEPENDS ON CERTAIN CONDITION THAT MIGHT BE INTERNAL OR EXTERNAL TO THE SUBJECT IN QUESTION

WHY ARE PEOPLE PUNISHED?

DETERRENCE
PREVENT OTHER ILLEGAL ACTIONS, IN ORDER TO WORK IT MUST BE ADRESSED TO PEOPLE WHO CAN ACTUALLY BE DETERRED

RETRIBUTION
THE PUNISHMENT SHOULD BE PROPORTIONED TO THE CRIME. SEVERE ENOUGH TO NOT COMMITT SIMILAR CRIMES BUT NOT TO BE UNJUST

MOST CURRENT LEGAL SYSTEM EMPLOY A MIXTURE OF THE TWO

- IN ORDER FOR A CRIME TO OCCUR WE NEED TWO COMPONENTS

CAN BE DONE BY ONE AI

FACTUAL COMPONENT (ACTUS REUS)
THE ACTION/OMISSION QUESTION ABOUT WHAT HAPPENED
↓
CONCERNS THE PHYSICAL ASPECTS OF THE POTENTIALLY CRIMINAL ACTION

? CAN BE DONE BY AN AI? AI DO NOT HAVE A MIND

MENTAL COMPONENT (MENS REA) → IN GENERAL IS A NECESSARY ELEMENT FOR ATTRIBUTION OF CRIMINAL RESPONSABILITY
DID THE AGENT ACT/OMIT IN RECKLESS OR MALICIOUS MANNER?
↓
FOCUSES ON WHETHER OR NOT THE CRIME WAS DONE WITH A GUILTY MIND

AI CAN IN PRINCIPLE FULFILL THOSE REQUIREMENTS (CLACIOIA-SARTOR)

COGNITION
AGENT AWARENESS OF REALITY GAINED WITH PERCEPTION, CONCEPTUALIZATION AND KNOWLEDGE

VOLITION
AMOUNT OF INTENTION WILLING TO DO SOMETHING

? CAN AN AI HAVE AN INTENTION TO PERFORM AN ACT?

- DIFFERENT SCHOOL OF THOUGHT
  - THERE IS SOME EVIDENCE OF INTENT WHEN AN AGENT CAN FORESEE THE PROBABLE OUTCOME OF AN ACTION
  - IF AN AGENT IS FREE TO CHOOSE AMONG DIFFERENT ACTIONS AND OPTS FOR ONE THIS PROVIDES EVIDENCE OF INTENT
  - MENS REA CAN BE ATTRIBUTED AT LEAST TO SOME AI SYSTEM
  - NO HUMAN CAN BE ATTRIBUTE OF CRIMINAL RESPONSABILITY
- RESPONSABILITY GAP THAT WE NEED TO FIX

TWO STRATEGIES FOR LAW IN AI

WHAT TO DO WHEN NEW TECHNOLOGY IS ALREADY ENTERING INTO OUR DAILY LIFE AND IS STILL NOT REGULATED?

- **LEGISLATIVE ROUTE (CREATIVE APPROACH):** YOU CREATE NEW NORMS AND NEW LEGAL CATEGORIES
  - SUPPOSE THAT THESE ARTIFICIAL AGENTS HAVE SOME SORT OF LEGAL PERSONHOOD → IMPOSSIBLE TO GIVE LEGAL PERSONHOOD TO THOSE AGENTS, NO MOMENT OF GENUINE AUTONOMY
  - THEIR POTENTIAL CREATIVITY MAKE THEM POSSESS LEGAL PERSOONHOOD (CAN BE DIFFICULT TO THE PROGRAMMERS TO PREDICT ALL THE OUTCOME) ← BUT
  - **THE EU AI ACT** →
    - COMPLITELY ABANDON THE IDEA OF RESPONSIBILITY BASED ON LEGAL PERSONHOOD FOR AI AGENTS AND INSTEAD ADOPT A RISK MANAGEMENT APPROACH
    - UNIFIED AND HARMONIZED FRAMEWORK FOR REGULATING AI IS INSIDE THE ACT ☺
    - EU ONE OF LARGEST AND MORE STRICT, WILL PROBABLY BE APPLIED IN THE REST OF THE WORLD ☺

  LIMITATIONS
  - IGNORES THE DEBATES ON THE POTENTIAL LEGAL PERSONHOOD OF AI AGENTS AND THE DIFFERENT NOTIONS OF RESPONSIBILITY THAT MIGTH BE USED FOR CONSTRAINING AI AGENTS
  - NEED TO DEFINE DIFFERENT LEVEL OF RISK
  - IF THE AGENTS ARE REALLY AUTONOMUS WHAT LEVEL OF RISK CAN COVER EVERY POSSIBILITY?
- **INTERPRETATIVE ROUTE (CONSERVATIVE APPROACH):** CLASSIFY THE NEW PHENOMENA IN EXISTING LAW
  - ASK IF AIs SHOULD OR NOT BE GRANTED MORE RIGTHS AND RESPONSIBILITIES AS THEIR AUTONOMY INCREASES. WE HAVE EXTENDED MORE RIGHTS TO SOME AGENTS WHICH PREVIOUSLY DID NOT HAVE THEM — ACQUIRING MORE RIGHTS DOES NOT ENTAIL THAT ONE NECESSARILY ACQUIRES MORE RESPONSIBILITY

    TREAT AI AS **AMBASSADORS**, THEY HAVE A CERTAIN SPHERE OF AUTONOMY BUT THEY DO NOT THEIR OWN WILL OR INTENTIONS. HAVE FREEDOM IN EXECUTING MISSIONS BUT THEY DO NOT HAVE MISSIONS ON THEIR OWN

    TREAT AI AS **SLAVES UNDER CLASSICAL ROMAN LAW**. CAN ACT ON THEIR OWN VOLITION BUT THEY ARE PROPERTY OF THE OWNER. GRANTED A CERTAIN LEVEL OF AUTONOMY

    LIMITATIONS
    - MANY AGENTS ARE NOT PHYSICALLY LOCATED: DIFFICULT TO DECIDE WHICH JURISTICTION TO APPLY
    - HARD TO DETERMINE ULTIMATELY WHO IS IN CHARGE OF THEM

    WHO IS RESPONSABLE?
    - **VICARIOUS RESPONSABILITY:** RESPONSABILITY SOMEONE HAVE FOR THE ACTIONS PERFORMED BY SOMEONE ELSE (A KID BREAKING A WINDOW, THE PARENT IS RESPONSABLE)
      ↓
      AI AGENTS DO NOT HAVE A LEGAL PERSONHOOD, THEY ARE TREATED SIMILARLY TO CHILDRENDS. EITHER THE OWNER, THE PROGRAMMER OR USER OF THE AI ARE RESPONSABLE FOR THE DAMAGE DOWN BY THE AI
    - **STRICT LIABILITY:** YOU ARE RESPONSABLE FOR THE CONSEQUENCES OF YOUR ACTIONS

THREATS AND TYPES OF AI CRIMES

TWO POSSIBLE STRATEGY

- THE POSSIBILITY OF CRIMES COMMITTED BY AN AI SYSTEM CAN THEMSELVES BE SUBJECT TO CRIMINAL LAW: AI SYSTEMS ARE NOT TO BE TREATED AS AGENTS OR ADRESSES OF THE CRIMINAL LAW BECAUSE ONLY HUMANS CAN BE SUBJECTS TO NORMS AND CRIMINAL LAW
- CONSIDER THE AI SYSTEM THEMSELVES AS SUBJECTS OF CRIMINAL LAW: AI CAN BE TREAT SIMILARLY TO HOW HUMANS ARE TREATED FOR SIMILAR VIOLATIONS

  FAROLDI ↓ (CLACIOIA-SARTOR)
  - THIS REQUIRES THAT AI SYSTEMS HAVE PERSONHOOD
    - NOT REQUIRED AI TO HAVE PERSONHOOD
      - KELSEN: "THE PERSON IS UNDERSTOOD AS THE TOTALITY OF RIGTHS AND OBLIGATIONS WHICH HAVE THE BEHAVIOUR OF A HUMAN BEING AS ITS CONTENT"
    - TO SAY THAT AN ENTITY IS A PERSON IS THE SAME AS THE ENTITY IS THE BEARER OF RIGTHS AND DUTIES
    - TWO CONCEPTION FOR LEGAL PERSONHOOD:
      - **THIN CONCEPTION:** AN ENTITY IS CONSIDERED A PERSON JUST IN CASE THAT THERE IS AT LEAST ONE NORM ADRESSING THE BEHAVIOUR OF THAT ENTITY PERHAPS ATTRIBUTING TO ITS RIGTHS OR DUTIES
      - **THICK CONCEPTION:** THE BEHAVIOUR OF THAT ENTITY MUST BE ADRESSED BY A SET OF NORMS CORRESPONDING TO THE NORMS GENERALLY APPLICABLE TO HUMANS

- WHAT ARE THE FUNDAMENTALLY UNIQUE AND PLAUSIBLE THREATS POSED BY AI CRIMES
- WHAT SOLUTIONS ARE AVAILABLE FOR DEALING WITH THEM?

↓

- FLORIDI: FOUR CAUSES FOR (RISK) CONCERN WITH REGARDS TO AI SYSTEMS

1) EMERGENT: AI DESIGNED IN CERTAIN WAY BUT WHEN IS "IN THE WILD" WILL PERFORM IN UNPREDICTED WAY. MAY ACT IN MORE SOPHISTICATED WAY THEN EXPECTED, CAN LEAD TO PROBLEMATIC STATES OF AFFAIRS

2) LIABILITY (GAP): (WHO IS RESPONSABLE?) THREAT THAT AI MODELS COULD UNDERMINE EXISTING LIABILITY MODELS

3) MONITORING: 3 PROBLEMS
- ATTRIBUTION: IDENTIFY WHO OR WHAT IS RESPONSIBLE FOR THE ACTIONS OR DECISIONS OF AN AI SYSTEM
- FEASIBILITY: THE PRACTICAL CHALLENGES OF OBSERVING OR INTERVENING IN THE DECISION MAKING PROCESS OF AI SYSTEMS
- CROSS-SYSTEM ACTIONS: THE POTENTIAL FOR AI SYSTEM TO INTERACT AND INFLUENCE EACH OTHER IN WAYS THAT ARE DIFFICULT TO PREDICT AND CONTROL

4) PYSCHOLOGY: THREAT OF AN AI AFFECTING THE MENTAL STATES OF INDIVIDUALS AND FACILITATES OR CAUSES CRIMES

- TYPES OF CRIMES THAT COULD BE PERFORMED BY AN AI:
  - FINANCIAL CRIMES
  - DRUG CRIMES
  - BOTS FULFILLING ILLEGAL ACTIONS
  - ADVANCED HUMAN-COMPUTER INTERACTION CAN BE USED TO PROMOTE SEXUAL OBJECTIFICATION, SEXUALIZED ABUSE/VIOLENCE
- THEFT, FRAUD: GATHER PERSONAL DATA AND SIMULATE PERSON
- ARTIFICIAL AGENTS CREATES NEW AND UNIQUE WAYS TO COMMIT EVERYDAY CRIMES OF A VARIETY OF NATURES

AVOID HARM

HOW TO MAKE A SYSTEM AVOID INNOCENTLY CAUSING HARM?

1) RESTRIC OR LIMIT THE SYSTEM IN WHAT IT DOES
   - LIMITING FUNCTIONAL CAPACITIES IN THE SPHERE OF ACTION
   - LIMITING IN THEIR DEPLOYMENT
2) ENDOW ARTIFICIAL AGENTS WITH SUPERIOR COGNITIVE CAPACITY SO THAT THEY CAN FIGURE OUT THE UNINTENDED EFFECTS OF THEIR ACTIONS
   - SEEMS UTOPIAN
   - SEEMS TO ASSUME THAT THE UNIVERSE IS COMPLETELY DETERMINISTIC AND EVERY COURSE OF ACTION CAN BE PREDICTED
   - IS PLAUSIBLE THAT ADVANCED AI AGENTS CAN FIGURE OUT MORE EFFECTS THAT A HUMAN MIGTH BE ABLE TO

- EMERGENT BEHAVIOURS:
  - LIMIT THE AGENT'S AUTONOMY OR ITS DEPLYMENT. THIS PREVENT TO REACH UNPREDICTED BEHAVIOUR
  - BUILD NORMATIVE AGENTS REQUIRING DEVELOPERS TO ENSURE THAT ARTIFICIAL AGENTS HAVE RUNTIME LEGAL COMPLIANCE LAYERS FIRMY IN PLACE BEFORE SUCH SYSTEM ARE DEPLOYED
- DISTINCTION BETWEEN SANCTIONS AND PRICE
  - A LAW MUST NOT HAVE A PRICE TAG BUT A FINE SHOULD DEPENDS ALSO ON THE PERPETUATOR
  - AI AGENT MIGTH BE INCAPABLE OF BEING RESPONSIVE TO MORAL OR OTHER NORMATIVE REASONS
  - (LACROIA AND SARTOR): AI SYSTEMS NEEDS TO BE RESPONSIVE TO BOTH MORAL AND LEGAL REASONS THAT IS THEY MUST POSSESSES NORMATIVE ARCHITECTURE

↓

THE CAPACITY TO TAKE VALUES AND NORMS INTO ACCOUNT

- THREE CAPACITY
  1) THE SYSTEM'S ABILITY TO GAIN OR HAVE AWARENESS OF ITS CONDUCT AND OF THE RESULTING EFFECTS OF IT
  2) THE SYSTEM'S ABILITY TO IDENTIFY AND UNDERSTAND THE NORMS THAT APPLY TO IT, THE SANCTIONS THAT CORRESPONDS TO NON-COMPLIANCE AND THE IMPACT THAT SANCTIONS HAVE ON THE AI AGENTS INTERESTS
  3) THE SYSTEM MORAL MOTIVATION TO COMPLY WHICH MIGTH BE DESIGNED FOR OR TRAINED IN A VARIETY OF WAY ?

HOW TO MAKE A SYSTEM AVOID INTENTIONALLY OR RECKLESSY CAUSING HARM?

- ENSURE THAT THE SYSTEM DO NOT ADOPT MALICIOUS ATTITUDES

1) PROVIDE APPROPRIATE DISINCENTIVES FOR MALICIOUS ACTIONS TO EITHER DEVELOPERS OR USERS
   - THE USER SPECIFY CERTAIN BEHAVIOURS TO AVOID AND IMPOSES NEGATIVE CONSEQUENCES WHEN THE BEHAVIOUR APPEARS
   - PUNISHED-BASED APPROACH
2) ENDOWING THE SYSTEM ITSELF WITH A NORMATIVE ARCHITECTURE ?
   - BUILD A SYSTEM THAT CAN RESPONDS TO NORMS IN A GENUINE WAY
   - AIM TO HAVE RESPONSES TO VALUES AND NORMS WHICH ARE ROOTED IN THE NORMATIVE, RATHER THAT STATISTICAL SIGNIFICANCE OF THOSE VALUES AND NORMS

LIABILITY

- DIRECT LIABILITY: CONSIDER AI AS POTENTIAL EQUAL DECISION-MAKER. IN CURRENT LEGISLATION CAN NOT BE CONSIDERED LEGALLY LIABLE IN THEIR OWN CAPACITY
- PERPETRATION BY ANOTHER: AI TREATED AS INSTRUMENT FOR CRIME AND NOT AS PROPER AGENTS, GUILTY WHO WANTED THE CRIME. IN COURT PROGRAMMER/ MANUFACTURES AND/OR USERS.
- COMMAND RESPONSIBILITY: KNOWLEDGE AS THE STANDARD FOR MENS REA. GUILTY WHO KNEW AND DID NOTHING TO STOP THE CRIME TO BE COMMITTED (USEFUL IF THERE IS A CLEAN CHAIN OF COMMAND, LIKE MILITARY)
- ? NATURAL PROBABLE CONSEQUENCE LIABILITY: NEGLIGENCE OR RECKLENESS AS A STANDARD OF MENS REA. ARM AS NATURAL CONSEQUENCE OF THEIR ACTION (EVEN IF THEY DID NOT INTENDED IT)
  - NOT CLEAR WHICH PROGRAMMER SHOULD BE CONSIDERED TO BE GUILITY

ALGORITHM TRASPARENCY AND DEEP LEARNING

- GDPR PROHIBITS AUTOMATED DECISIONS THAT HAVE CONSEQUENCES FOR INDIVIDUALS AND ESTABLISHES A RIGTH TO MEANINGFUL INFORMATION ABOUT THE LOGIC ADOPTED IN THESE PROCEDURES

- REASONING SYSTEM: DESIGNED TO PERFORM LOGICAL REASONING/DEDUCTION. SYMBOLIC REPRESENTATION OF SOURCE

  ↓

  PROBLEM OF TRANSPARENCY AND THE LOGIC OF OPERATION DO NOT ARISE

- LEARNING SYSTEM: DESIGNED TO LEARN FROM DATA USE IN ABSENCE OF SYMBOLIC ORGANIZATION OF SOURCE DATA

  ↓

  PROBLEMS IN TERMS OF TRANSPARENCY AND LOGIC OF OPERATION

MAKING AN ALGORITHM TRASPARENT

3 DIFFERENT TERMS

- INTENTIONAL OPACITY: THE ALGORITHM IS VOLOUNTARY NOT PUBLIC TO MAINTAIN A COMPETITIVE OR ECONOMIC ADVANTAGE.
  - CAN BE FIXED BY PUBLISHING THE ALGORITHM
- COGNITIVE OPACITY: EVEN IF TRANSPARENT CAN NOT BE INTERPRETED BY MOST USERS. NEED INTERPRETERS OR MASSIVE INVESTMENT IN EDUCATION
- INHERENT OR ESSENTIAL OPACITY: THE ALGORITHM MAKE DECISIONS TO NOT BE INTERPRETED BY HUMANS

MAYBE EU COMMISSION DON'T WANT TO OVERREGULATE A BOOMING SECTOR

TWO BROAD CRITICISM OF THE PROPOSED EU AI ACT

1) THE PROPOSAL DOES NOT ADRESS THE FUTURE DEVELOPMENTS OF GENERAL INTELLIGENCE SYSTEM AND IT IS ILL-EQUIPPED TO DEAL WITH MORE GENERAL EXISTING SYSTEM LIKE GPT-3
2) THE PROPOSAL FAILS ABOUT TRANSPARENCY BECAUSE IS PROBLEMATIC TO SPELL TRANSPARENCY OUT IN A CONCRETE AND USABLE WAY

CONTROL PROBLEM: HOW CAN WE BE SURE THAT GENERAL-INTELLIGENCE (OR SUPERINTELLIGENCE) AGENTS DO NOT TAKE CONTROL OVER THE WORLD
- EU AI ACT
  - FAILS TO PROPERLY TAKE INTO ACCOUNT THE CONTROL PROBLEM DUE TO POTENTIAL GENERAL AI
  - FAILS TO ACCOUNT FOR EXISTING NON-SPECIFIC SYSTEM (GPT-3)

SOLUTIONS FOR THE CONTROL PROBLEM:
- ALIGNMENT: SUPERINTELLIGENCE AI MUST BE PROGRAMMED TO ALIGN WITH HUMAN VALUES
  - GENERAL AGREEMENT THAT ALIGNMENT HAS TO BE BUILD BEFORE A SINGULARITY IS REACHED
  - EU AI ACT FOLLOW AND DO NOT SHAPE TECHNOLOGICAL DEVELOPMENT. GRAVE MISTAKE. GENERAL IDEA THAT WHEN SINGULARITY WILL BE REACHED IT WILL BE TO LATE TO DO ANYTHING ABOUT CONTROL

- AI SYSTEM AT HIGH-RISK: APPLICATION IN A FIELD THAT IS CONSIDERED PARTICULARLY RELEVANT OR WORTH EXTRA CARE
  - RISK IN CLASSIFY A LOW-LEVEL SPECIFIC AI (CLASSIFICATION AS HIGH RISK, AND A SUPERINTELLIGENCE WITH NO CONCRETE APPLICATION AS LOW-RISK
  - EXCLUDE CURRENTLY EXISTING SYSTEM THAT DO NOT HAVE A SPECIFIC CONCRETE INTENDED APPLICATION BUT ARE FAIRLY GENERAL

TRASPARANCY IN EU ACT + SEE BEFORE 3 OPACITY

- TRASPARENCY IS NOT A CLEAR CONCEPT, IT IS ABOUT KNOWING HOW A SYSTEM WORKS
- TRASPARENCY IN THE EU ACT IS NOT ALIGNED WITH THE REST OF LITERATURE. IN EU ACT IS THE RIGTH TO BE TOLD THAT ONE IS INTERACTING WITH AN AI SYSTEM
- REQUIRES SOME FORMS OF CONTROL AND TRASPARENCY FOR HIGH-RISK SYSTEMS

- **INTERPRETABILITY**
  - **LIPTON**: INTERPRETABILITY OF MODELS FALLS INTO TWO BROAD CATEGORIES
    - STRICTO SENSUS: MATTERS THE MECHANISM BY WHICH THE MODEL WORKS
    - POST HOC: MATTERS TO EXTRACT INFORMATION FROM THE MODELS TO CLARIFY WHAT EXACTLY THEY LEARNED
  - **KRISHNAN**: CONCEPT OF INTERPRETABILITY IS VAGUE. ALSO IT IS A MEAN TO ARCHIVE OTHER ENDS (NON-DISCRIMINATION,...), WE NEED TO FOCUS ON THE TRUE ENDS WE WANT TO ARCHIEVE

PHYLOSOPHICAL CONSIDERATIONS ON THE STATUS OF SUPERINTELLIGENT ARTIFICIAL AGENTS

- IDEA THAT ARTIFICIAL AGENTS CAN ENTER FULLY INTO THE ATTRIBUTION OF RESPONSIBILITY (MORAL, CRIMINAL)
- PROBLEMS WITH ETHICAL AND LEGAL IMPLICATIONS DUE TO THE FACT THAT AI SYSTEMS WORKS TOO WELL (DEEPFAKE, REPLACE LOW-SKILLED JOBS,...)
- INTELLIGENT AGENTS INTERFERE WITH DIFFERENT PARTS OF SOCIETY ON A DAILY BASIS
- WHAT WOULD BE THE STATUS (LEGAL, ETHICAL AND NORMATIVE) OF A SUPERINTELLIGENCE AGENT?
  - HOW SHOULD WE TREAT SUPERINTELLIGENT AGENTS?
  - HOW CAN WE BE SURE THAT THEY TREAT US IN THE WAY WE WANT TO BE TREATED?
- WHAT SHOULD BE THEIR STATUS?
- CONTROL PROBLEM
  - ALIGNMENT
  - LIMIT THE CAPACITIES OF AI SYSTEMS BY ISOLATING IT FROM THE OUTSIDE WORLD
  - INCREASE THE CAPABILITIES OF HUMANS TO BE ON PAR WITH SUPERINTELLIGENT SYSTEM

- **DEFINITION OF INTELLIGENCE**: IS PROBLEMATIC, MUST BE DISTINGUED FROM CONSCIOUSNESS AND AWARENESS.
- TURING IMITATION GAME
- SEARLE CHINEESE ROOM
- INTELLIGENCE FACTOR (g)

ISSUES OF NON-HUMAN AGENT WITH INTELLIGENCE SIMILAR/SUPERIOR TO HUMANS

TECHNICAL ISSUES

- NORMATIVE SYSTEMS ARE WRITTEN IN NATURAL LANGUAGE AND THEREFORE A TRANSLATION IS NEEDED
- NEED AT LEAST SECOND-ORDER LOGIC TO TRANSLATE NORMATIVE SYSTEMS IN FORMAL LANGUAGE (IT IS INCOMPLETE)
- STANDARD DEONTIC LOGIC IS
  - INADEQUATE TO RENDER ALL THE COMPLEXITIES OF NORMATIVE LANGUAGE
  - FULL OF PARADOXES AND PROBLEMS WITH COUNTER-INTUITIVE CONSEQUENCES

- **LIABILITY**:
  - ITALIAN REPUBLICAN CHARTER PRESCRIBES THAT CRIMINAL LIABILITY IS PERSONAL
  - NO STRICT LIABILITY IN THE CURRENT LEGAL SYSTEM
  - LIABILITY OF ENTITIES FOR ADMINISTRATIVE OFFENSES ARISING FROM CRIMES (AKA CRIMINAL LIABILITY FOR ENTITIES)
  - IT IS POSSIBLE FOR INTELLIGENT-NON-HUMAN AGENTS TO BE MORAL AGENTS
    - MORALITY RULES AND DECISION-MAKING PROCESSES ARE ENCODED (REGUS: PRE-PROGRAMMED) AB INITIO
    - AUTONOMOUS AGENTS ARE ABLE TO REASON MORALLY BY SELF-ISTRUCTION (BY EXTRACT MORAL PATTERNS FROM LARGE DATASETS)

- **OBJECTION, ALTERNATIVE PROPOSAL AND OPEN QUESTIONS**
  - ANY SANCTION ATTRIBUTABLE TO NON-HUMAN AUTONOMOUS AGENTS IS NOT SO MUCH A PENALTY AS A SECURITY MEASURE
    - HAVING ASCERTAINED THE SOCIAL DANGEROUNESS OF THE AUTONOMOUS NON-HUMAN AGENT, IT WOULD IN THE ABSENCE OF ITS OWN SUBJECTIVE ELEMENTS NOT BE IMPUTABLE AND THEREFORE NOT PUNISHABLE
  - INTRODUCE A SPECIFIC CRIMINAL LAW FOR SUPERINTELLIGENT AUTONOMOUS AGENTS
    - PRINCIPLE OF PERSONALITY OF LAW: INTERACTION BETWEEN DIFFERENT NATIONS
    - HYBRIDS NEEDS TO BE CAREFULLY CONSIDERATED
    - QUESTIONABLE TO JUSTIFY WHY IT SHOULD NOT BE THE SUPERINTELLIGENCE IMPOSE CERTAIN RULES TO US AND NOT VICE-VERSA

- **DELPHI** FOR AN AI TO BEHAVE ETHICALLY, TRAINING DATA SHOULD NOT BE BASED ON WIDESPREAD PRACTICES
  - ONE CAN TYPE QUESTIONS AND GET ANSWERS LIKE "IS OK" OR "YOU SHOULDN'T"
  - USE DESCRIPTIVE ETHICS: IS ABOUT ETHICAL BELIEFS OF A SOCIETY IN A GIVEN TIME. NORMATIVE/PRESCRIPTIVE ETHICS INSTEAD IS ABOUT WHAT IS RIGTH TO DO. HUMANS ARE IMPERFECT BASING ANSWERS ON WHAT PEOPLE SAYS IS RISKY
  - MISTAKE BETWEEN NORMS AND NORMAL:
    - NORMAL: THINGS THAT STATISTICALLY ARE NORMAL (SLAVERY IN ANCIENT GREEK) STICKING TO WHAT IS NORMAL TODAY WILL IMPEDE MORAL PROGRESS
  - HOW TO ASK QUESTIONS: ANSWERS DEPENDS A LOT ON THE GRAMMAR STRUCTURE OF THE QUESTION

HUMAN COMPATIBLE APPROACH

- ISSUE OF EMBEDDING HUMAN VALUES IN ARTIFICIAL GENERAL INTELLIGENCE (AGI). AS THEY BECOME MORE INTELLIGENT THEY COULD CHOOSE METHODS AND MEANS NOT ALIGNED WITH HUMANS
- **HUMAN COMPATIBLE APPROACH** (RUSSEL): INTELLIGENT AGENTS NOT REQUIRED TO MAXIMIZE A SIMPLE REWARD FUNCTION BUT MAXIMIZE THE REALIZATION OF HUMAN PREFERENCES WHICH ARE ESSENTIALY UNCERTAIN
- PREVENT AI CONTROL BY PUTTING HUMANS AT THE CENTER OF INTELLIGENT MACHINE DESIGN, AND DOES SO UNIQUELY BY POOLING EXPERTISE IN PHILOSOPHY, LOGIC AND COMPUTER SCIENCE
- AGI WILL BE REACHED WITHIN THE END OF THE CENTURY
- INSTRUCTED TO OPTIMIZE A REWARD FUNCTION (DONE BY DESCRIVING PRECISELY THE COMBINATION OF GOALS)
- **AI ALIGNMENT PROBLEM** (BOSTROM): MIGTH CHOOSE METHODS AND MEANS NOT ALIGNED WITH HUMANS

- TO REACH THE LATTER GOAL USE INVERSE REWARD FUNCTION (IRL). REWARD FUNCTION ARE NOT GIVEN BUT HAVE TO BE LEARNT
- OBSERVE HUMANS AND LEARN THEIR OBJECTIVE, AIMS AND VALUES
- 3 PROBLEMS
  1) YOU CAN OBSERVE (SOME) BEHAVIOUR, BUT BELIEFS AND VALUES ARE DIRECTLY UNOBSERVABLE
  2) MODELING OFTEN RELIES ON BEHAVIOR THAT IS ASSUMED TO BE OPTIMAL. SHOULD BE ABLE TO UNDERSTAND WHEN HUMANS MAKE MISTAKES IN ANY GIVEN TIME
  3) EVEN IF LEARN EVERYTHING AND START TO MAXIMISE UTILITY IT COULD LEAD TO MISALIGNMENT WITH HUMAN VALUES. UNCERTAINTY IS EPISTEMIC, NOT ONTOLOGICAL AND IS STATIONARY

NORMATIVE RISK APPROACH

POTENTIAL FOR HARM THAT ARISE FROM THE VERY EXISTENCE OF CERTAIN NORMS AND VALUES

NORMATIVE RISK AS A PROBABLE OR MORAL HARM:

- LIKELYHOOD OF DOING HARM THROUGH ACTION THAT VIOLATES NORMS OR VALUES. DEPENDS ON THE BACKGROUND NORMATIVE (MORAL) THEORY

↑ POTENTIAL FOR HARM AND NEGATIVE CONSEQUENCES THAT ARISE FROM ACTION THAT VIOLATE NORMS OR VALUES

NORMATIVE RISK AS NORMATIVE UNDERDETERMINACY

- VALUE GAPS: SITUATIONS THAT ARE NORMATIVELY INDIFFERENT. LACK OF CLEAR NORMS AND VALUES TO GUIDE DECISION MAKING BEHAVIOUR (UNDERDETERMINACY)
- VALUE GUTS: SITUATIONS THAT ARE NORMATIVELY QUALIFIED IN INCOMPATIBLE WAYS. CONFLICT OF NORMS OR VALUES THAT MAKE DIFFICULT TO DETERMINE THE RIGTH COURSE OF ACTION (OVERDETERMINACY)

NORMATIVE RISK AS NORM-RELATED EXISTENTIAL RISK. TWO MEANING OF "RELATED"

- CAUSATION: NORMS AND VALUES INCREASE THE PROBABILITY OF EXISTENTIAL RISK
- OMISSIONS: FAIL TO DECREASE THE PROBABILITY OF AN EXISTENTIAL RISK

CONSIST OF ANALYSING HOW EXISTING OR PROPOSED NORMS ENGENDER OR FAIL TO PREVENT FUTURE CATASTROPHIC EVENTS OR EXISTENTIAL RISK

STRATEGIES TO MITIGATE NORMATIVE RISK → IDENTIFICATION OF NORMATIVE RISK → ANALYSIS & QUANTIFICATION → ELABORATION OF MITIGATING STRATEGIES → IMPLEMENTATION
- LONG TERM
- SHORT TERM

3 CRITERIAS TO DECIDE WHICH EXISTENTIAL RISK TO TACKEL FIRST
- IMPORTANCE: HOW MUCH A SPECIFIC RISK CONTRIBUTES TO THE OVERALL RISK? WHAT IS THE VALUE OF MITIGATING IT?
- TRACTABILITY: HOW EASY IS TO MITIGATE OR SOLVE THE PROBLEMS
- NEGLECT: HOW MUCH ATTENTION AND RESOURCES ARE DEVOTED TO SOLVE IT

EU AI ACT: RISKS

1) UNACCETTABLE RISK: FORBIDEN ACTIVITIES
2) HIGH RISK: OBLIGATORY REQUIREMENTS AND PRE-CONFORMITY CHECKS
3) LOW RISK: TRASPARENCY REQUIRED
4) MINIMAL RISK: TRANSPARENCY REQUIRED

↳ EU AI ACT DOES NOT ENGAGE IN PROVIDING A GENERAL DEFINITION OF RISK NOR A GENERAL STRATEGY TO IDENTIFY RISKS. RISKS ARE DETERMINE WITH LISTS
- MANY ACTIVITY NOT-RISK PRONE WILL BE CLASSIFIED AS RISKY AND MAYBE BANNED
- MANY ACTIVITY NOT IN THE LIST BUT THAT ARE RISKY WILL BE ALLOWED

\+ • FAILS TO TAKE INTO ACCOUNT GENERAL INTELLIGENT SYSTEM (AND GPT-3)
- TAKE INTO ACCOUNT OBJECTIVES SPECIFIED BY HUMANS (IGNORE THE PROBLEM OF AN AGENT DESIGNING ITS OWN OBJECTIVES)

NORMATIVE RISK APPROACH: ONE HAVE TO INDIVIDUATE AND MITIGATE THE EVENTS THAT INCREASE THE PROBABILITY OF AN EXISTENTIAL RISK

- NORMATIVE RISK LAB: PERFORM THEORETICAL RESEARCH ON DEFINING AND MODELING NORMATIVE RISK, ELABORATE GENERAL MITIGATION STRATEGIES AND POTENTIALLY ELABORATE POLICY RECCOMENDATIONS
- NORMATIVE RISK CONSULTING:
  - PRIVATE COMPANIES: COMPLIENCE CHECK ON EXISTING LEGISLATIONS
  - GOVERNMENTS: PRELIMINARY CHECK ON PROPOSED LEGISLATIONS
- NORMATIVE RISK INDEPENDENT ADMINISTRATIVE AUTORITY: ANALYZE PREEMTIVELY PROPOSED LEGISLATION FOR NORMATIVE RISKS, PUBLISH GUIDELINES AND DO POSTERIOR CHECKS, WILL HAVE REAL POWER TO INTERVENE ON PUBLIC AND PRIVATE INITIATIVES (BOTH BEFORE AND AFTER)
  - INDIPENDENT FROM GOVERNMENT BUT STILL SUBJECT TO LEGISLATIVE AUTHORITY: POLITICAL PRESSURE CAN POTENTIALLY COMPROMISE ITS EFFECTIVINESS

DEFINITIONS

- WEAK AI: PROGRAMS THAT DO NOT EXPERIENCE CONSCIOUSNESS OR DO NOT HAVE A MIND IN THE SAME SENSE PEOPLE DO. IMITATE HUMAN TASK WITHOUT BEING CONSCIOUS.
- STRONG AI: ABILITY OF AN INTELLIGENT AGENT TO UNDERSTAND, FEEL OR THINK LIKE A HUMAN → FOR SEARLE IT HAVE CONSCIOUSNESS, INTENTIONALITY AND UNDERSTANDING
- NARROW AI: ABILITY OF AN INTELLIGENT AGENT TO LEARN AND PERFORM A SPECIFIC TASK OFTEN WITH AT LEAST HUMAN PROFICENCY
- GENERAL AI: ABILITY OF AN INTELLIGENT AGENT TO LEARN AND PERFORM ANY INTELLECTUAL TASK THAT A HUMAN BEING CAN
- SUPERINTELLIGENCE: A HYPOTHETICAL AGENT THAT WOULD POSSESS INTELLIGENCE FAR SURPASSING THAT OF THE BRIGHTEST AND MOST GIFTED HUMAN MIND
- SINGULARITY: A HYPOTHETICAL POINT IN TIME AT WHICH TECHNOLOGICAL GROWTH BECOMES UNCONTROLLABLE AND IRREVERSIBLE, RESULTING IN UNFORESEEABLE CHANGES TO HUMAN CIVILIZATION
- MISALIGNMENT (MINIMALIST): AN AI A IS MISALIGNED WITH A HUMAN H IF H WOULD WANT A NOT TO DO WHAT A IS TRYING TO DO
- ALIGNMENT (MAXIMALIST): AN AI THAT INCORPORATES VALUES AND BEHAVES MORALLY (OR LEGALLY)
- CONTROL PROBLEM: MAKE SURE THAT SUPERINTELLIGENCE AGENTS DO NOT TAKE CONTROL OVER US AND THE WORLD
- EXISTENTIAL RISK: AN EVENT THAT INCREASES THE PROBABILITY OF ERADICATING LIFE ON THE PLANET
- NORMATIVE RISK (REGULATORY SENSE): EVENTS THAT INCREASE OR FAIL TO DECREASE THE PROBABILITY OF AN EXISTENTIAL RISK
- ARTIFICIAL GENERAL INTELLIGENCE (AGI): TYPE OF AI ABLE TO PERFORM ANY INTELLECTUAL TASK A HUMAN BEING CAN
- ETHICS (= MORAL): IS ABOUT WHAT ONE SHOULD DO ALL THINGS CONSIDERED
- METAETHICS: STUDY THE NATURE OF MORALITY AND ETHICAL THOUGHT. ASKS QUESTIONS ABOUT THE NATURE OF MORAL JUDMENTS AND PRINCIPLES AND HOW THEY RELATE TO HUMAN BEHAVIOR
- APPLIED ETHICS: DEALS WITH PRACTICAL APPLICATION OF MORAL PRINCIPLES TO SPECIFIC ISSUES AND PROBLEMS. IT ADRESSES MORAL QUESTIONS THAT ARISES IN SPECIFIC FIELDS OR CONTERTS (MEDICAL ETHICS, BUSINESS ETHICS,...)
- ONTOLOGY: STUDY OF THE NATURE OF EXISTENCE OR BEING. IT DEALS WITH QUESTIONS ABOUT WHAT TYPE OF THINGS EXISTS AND WHAT THEIR FUNDAMENTAL NATURE IS
- ALGORITHM: PRECISE, FINITE PROCEDURE THAT SPECIFIES HOW TO OBTAIN OUTPUT DATA FROM INPUT DATA. DOES NOT REQUIRE CREATIVITY AND CAN BE EXECUTED AUTOMATICALLY
- ARTIFICIAL INTELLIGENCE: CONSISTS OF SYSTEM THAT FUNCTION BY DEVELOPING OR POSSESSING CHARACTERISTIC THAT, IN HUMANS, CORRESPOND TO HIGHER COGNITIVE FUNCTIONS, SUCH AS THINKING, PROBLEM SOLVING AND LEARNING
- AI EFFECT: IDEA THAT LOT OF TASK CONSIDERED IN THE PAST INTELLIGENT, WITH THE ARRIVAL OF AI NOW ARE NOT INTELLIGENT ANYMORE