JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

Enabling machines to reason about potential harms to humans

AAAI 2022 Spring Symposium Series

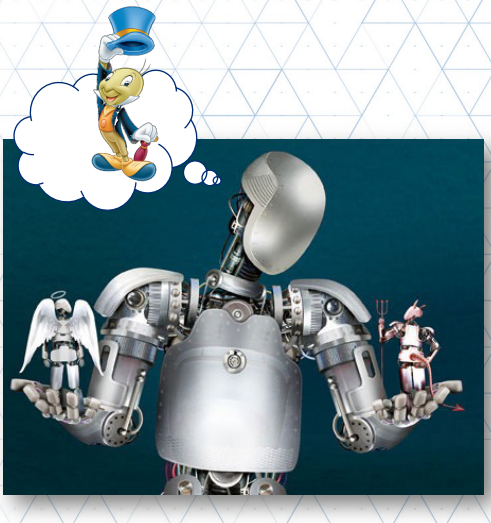
Putting AI in the Critical Loop: Assured Trust and Autonomy in Human-Machine Teams

(Approaches to Ethical Computing: Metrics for Measuring AI's Proficiency and Competency for Ethical Reasoning)

22 March 2022 1615-1700 EDT

AM Greenberg
Email: ariel.greenberg<at>jhuapl.edu
IPAish: Aw-ree-el
Pronouns: he | him | his

Senior Research Scientist
Intelligent Systems Center



1

Enabling machines to reason about potential harms to humans
Outline

- **Foundations**
  - **Ethics with respect to Artificial Agency**
- Perception
  - Moral-scene assessment (recap)
- Reasoning
  - Harm ontology
- Upcoming Publications

2

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

Foundations

Enabling machines to reason about potential harms to humans

State of the art of Human-Avatar IX

Desiderata for Intelligent Systems Technology

Collaboration toward AMA

Glossary

Artificial Ethical Agency vs Ethical use of artificial agency

Granting autonomy to artificial agency

Where is judgement happening in HMT?

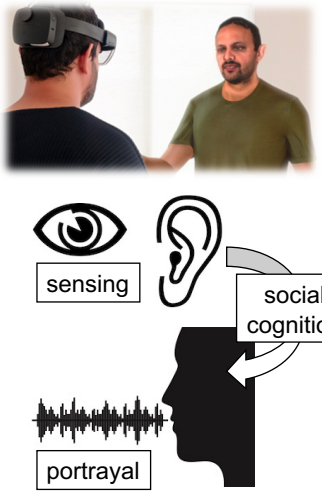
Moral reasoning & Trustworthiness

3

State of the art of Human-Avatar IX

Technology is outpacing science

- Sensing of interlocutor is OK
  - Computer vision
  - Standoff psychophysiology
  - Natural language processing
- Portrayal to interlocutor is excellent, in particular for virtual agents
  - Computer graphics rendering of face and body
  - Voice production
- But profound deficits in social cognition, dramatically incommensurate with progress in sensing and portrayal
  - Superficially accurate, but hollow inside: expectations diverge
  - Humans are suckers for this, so impressions are easily manipulated

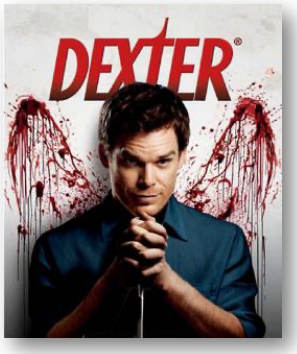


We have a name for this condition when it occurs in humans …

4

Psychopathy!

when expression is generated to manipulate, rather than to reflect internal processes



Let's not do this.

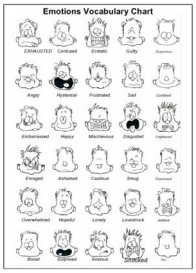
5

Desiderata for Intelligent Systems Technology

Machines that operate in accordance with, and in service of, human values

Socially competent
Instrumental convergence thesis

Affect sensitive



Values aligned
Orthogonality thesis

Let's do this instead.

6

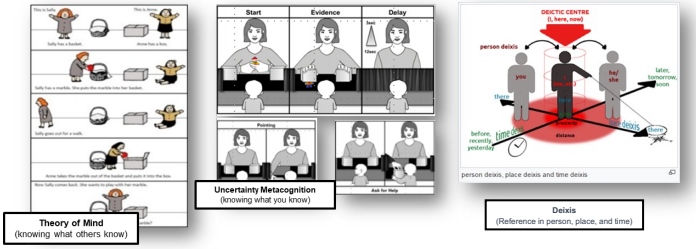
Social Competence

Desiderata for IST

Instrumental convergence thesis

Projects in Prosociality

- ~SCARAB: Social Competence Assessment of Robotic Autonomous Behavior
- PARTNER/PARTI: Platforms for Assessing Relationships: Trust with Near Ecologically-valid Risk, and Team Interaction*
- L2RM: Learning to read minds



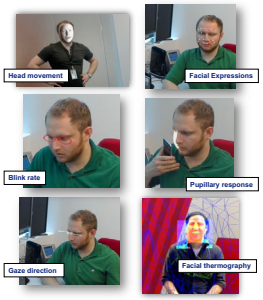
7

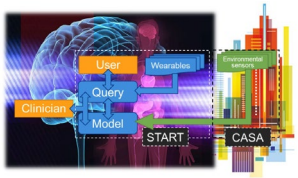
Affect Sensitivity

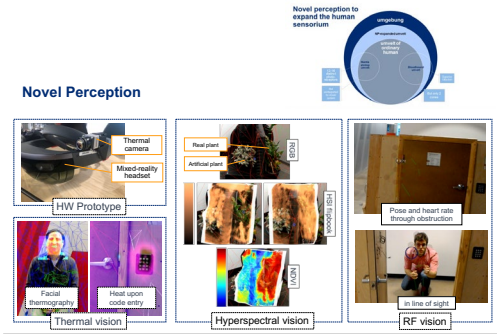
Desiderata for IST

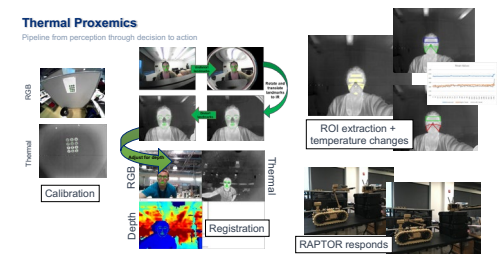
Human state estimation

- Social Prosthetics* & Novel Perception*
- CAPTIVA: Cardio- And Pulmonary Trends and Inference from Video Analytics
- Thermal Proxemics
- START: State-Triggered Assessment & Regulation Tool & ~CASA: Circumstance and Ambience Sensing for the Armamentarium









8

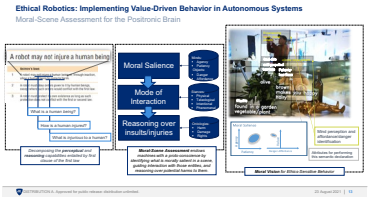
Values Alignment

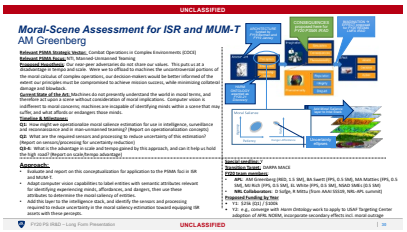
Desiderata for IST

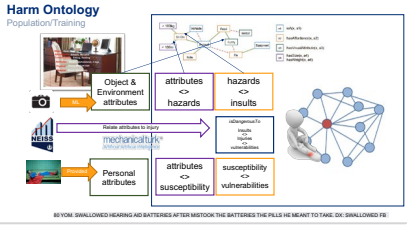
Orthogonality thesis

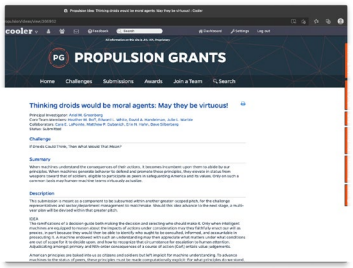
Human values encoding

- Ethical Robotics: Implementing Value-Driven Behavior in Autonomous Systems
- Moral-Scene Assessment for Intelligent Systems *
- Enabling autonomous systems to reason over potential harms to humans
- ~ MSA for ISR & Artificial JAG paralegal









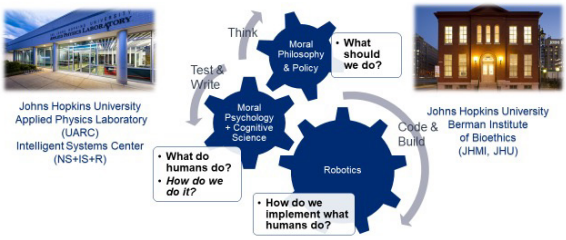
9

Machine Ethics @ JHU

APL Intelligent Systems Center[1] & Berman Institute of Bioethics[2]

Team: Ariel Greenberg[1], Travis Rieder[2], David Handelman[1], Debra Mathews[2]

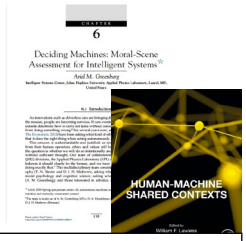
Story: Five years ago this month, we began our collaboration to conceive of an artificial proto-conscience that could see the world in moral terms, and then generate behaviors that serve to promote held values.



Results:

- JHU Practical Ethics grant (first APL awardee)
- JHU Discovery award (first Berman awardee)

- Many presentations, publications, and press
  - Two selected major publications shown to the right
  - *Received APL Outstanding Special Publication Award



*Deciding Machines: Moral-Scene Assessment for Intelligent Systems**

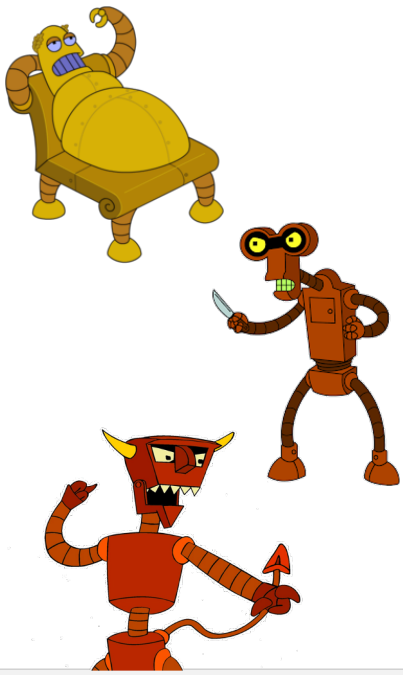
Artificial Intelligence in Service of Human Needs: Pragmatic First Steps Toward an Ethics for Semi-Autonomous Agents

10

Provincial Glossary

Ethics with respect to Artificial Agency

- ***Artificial agency***, meant to precipitate out a key element common to artificial intelligence and autonomous systems
  - ***Agent***: can instantiate intentional mental states capable of performing actions
    - ***Moral agency:*** an agent having the capacities for making free choices, deliberating about what one ought to do, and understanding and applying moral rules correctly in the paradigm cases
  - ***Autonomous systems***: Decision-making technology both capable and worthy of being granted some degree of independence from human control
    - ***Trust in Autonomy***: Granting autonomy to artificial agency necessarily incurs risk/vulnerability, and the willingness to accept that risk/vulnerability is what we call *trust*
- ***with respect to***, meant to capture two distinct relationships of ethics to artificial agency:
  - of (as object): Ethical Use of Artificial Agency
  - **for (as subject): Artificial Ethical Agency (previously AMA)**



11

Artificial Ethical Agency vs Ethical Use of Artificial Agency

Foundations

Artificial Ethical Agency

Regulating machines

Endogenous/organic/on-board

Principles pertain to artificial agent

**Commensurate capability and judgement*

e.g., non-maleficence, beneficence

Principles

Transparency
Justice and fairness
Non-maleficence
Responsibility
Privacy
Beneficence
Freedom and autonomy
Trust
Sustainability
Dignity
Solidarity

Ethical use of artificial agency

Governing humans

Extrinsic constraints

Principles pertain to designers and users

e.g., fairness, bias, privacy

Unitary apparatus for natural/motivated systems, differentiable for artificial systems

USG: IC, DOD, CIV (P)

ASILOMAR (P)

Rome Call (P)

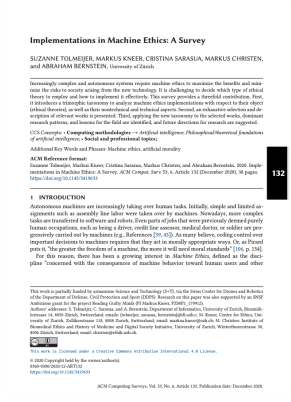
Harvard Report (D)

Montreal Report (D)

12

Artificial Ethical Agency vs Ethical Use of Artificial Agency

In lit, taxonomized


Implementations in Machine Ethics: A Survey

SUZANNE TOLMEIJER, MARKUS KNEER, CRISTINA SARASUA, MARKUS CHRISTEN, and ABRAHAM BERNSTEIN, University of Zürich

Increasingly complex and autonomous systems require machine ethics to maximize the benefits and minimize the risks to society arising from the new technology. It is challenging to decide which type of ethical theory to employ and how to implement it effectively. This survey provides a threefold contribution. First, it introduces a typological framework for machine ethics implementations with respect to their object (ethical theories), as well as their nonfunctional and technical aspects. Second, an exhaustive selection and description of relevant works is presented. Third, applying the new taxonomy to the selected works, dominant research patterns, and lessons for the field are identified, and future directions for research are suggested.

CCS Concepts: • Computing methodologies → Artificial intelligence; Philosophical/theoretical foundations of artificial intelligence; • Social and professional topics;

Additional Key Words and Phrases: Machine ethics, artificial morality

ACM Reference format:
Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2020. Implementations in Machine Ethics: A Survey. ACM Comput. Surv. 53, 6, Article 132 (December 2020), 38 pages.
https://doi.org/10.1145/3419633

132

1 INTRODUCTION

Autonomous machines are increasingly taking over human tasks. Initially, simple and limited assignments such as assembly line labor were taken over by machines. Nowadays, more complex tasks are transferred to software and robots. Even parts of jobs that were previously deemed purely human occupations, such as being a driver, credit line assessor, medical doctor, or soldier are progressively carried out by machines (e.g., References [39, 65]). As many believe, ceding control over important decisions to machines requires that they act in morally appropriate ways. Or, as Picard puts it, "the greater the freedom of a machine, the more it will need moral standards" [104, p. 134].

For this reason, there has been a growing interest in Machine Ethics, defined as the discipline "concerned with the consequences of machine behavior toward human users and other

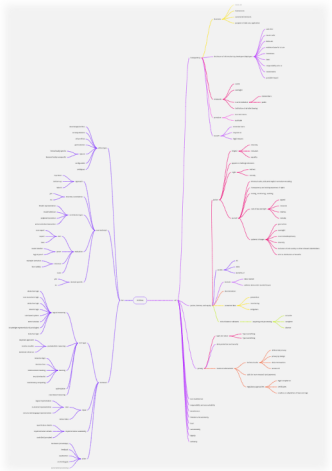
This work is partially funded by armasuisse Science and Technology (S+T), via the Swiss Center for Drones and Robotics of the Department of Defense, Civil Protection and Sport (DDPS). Research on this paper was also supported by an SNSF Ambizione grant for the project Reading Guilty Minds (PI Markus Kneer, PZ00P1_179912).

Authors' addresses: S. Tolmeijer, C. Sarasua, and A. Bernstein, Department of Informatics, University of Zurich, Binzmühlestrasse 14, 8050 Zürich, Switzerland; emails: {tolmeijer, sarasua, bernstein}@ifi.uzh.ch; M. Kneer, Centre for Ethics, University of Zurich, Zollikerstrasse 118, 8008 Zürich, Switzerland; email: markus.kneer@uzh.ch; M. Christen, Institute of Biomedical Ethics and History of Medicine and Digital Society Initiative, University of Zurich, Winterthurerstrasse 30, 8006 Zürich, Switzerland; email: christen@ibme.uzh.ch.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2020 Copyright held by the owner/author(s).
0360-0300/2020/12-ART132
https://doi.org/10.1145/3419633

ACM Computing Surveys, Vol. 53, No. 6, Article 132. Publication date: December 2020.




nature machine intelligence PERSPECTIVE

The global landscape of AI ethics guidelines

Anna Jobin, Marcello Ienca and Effy Vayena*

In the past five years, private companies, research institutions and public sector organizations have issued principles and guidelines for ethical artificial intelligence (AI). However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. To investigate whether a global agreement on these questions is emerging, we mapped and analysed the current corpus of principles and guidelines on ethical AI. Our results reveal a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy), with substantive divergence in relation to how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to, and how they should be implemented. Our findings highlight the importance of integrating guideline-development efforts with substantive ethical analysis and adequate implementation strategies.

(for, as subjects)

(of, as objects)

13

Granting autonomy to artificial agency is special compared to use of AI

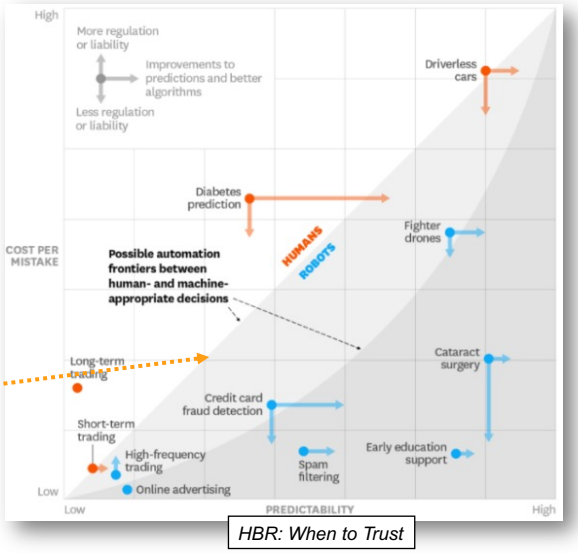
Foundations

Special features include:

- Independent judgement and decision making
- Dual learning: of "how" and of "ought"
- Learning, Opaque lessons
- Emergent behavior
- Unexpected conditions
- Need for social cognition
- Human vulnerability to moral consequences of machine action

Of course, we'd want humans to be the ones drawing the line, but in practice, it may be by machines, or blurred.

We need to be able to trust this too, (or especially)!

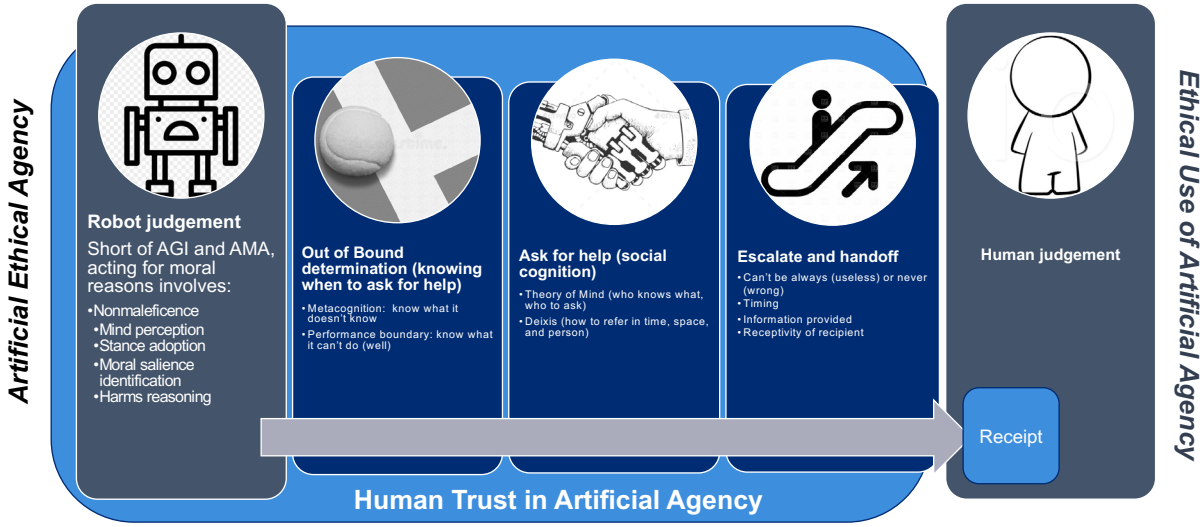


HBR: When to Trust

14

Ethics wrt Artificial Agency

Where is judgement happening?

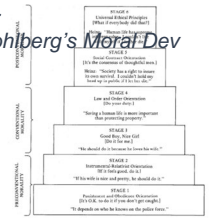


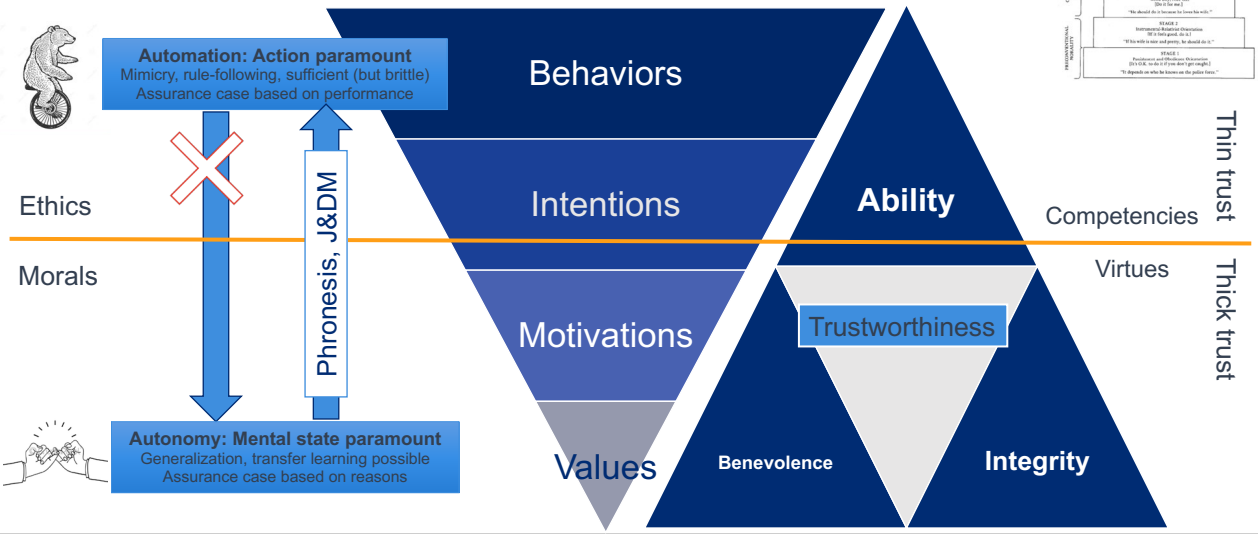
15

Ethics wrt Artificial Agency

Moral reasoning & Trustworthiness

Next step:
Cf with Kohlberg's Moral Dev

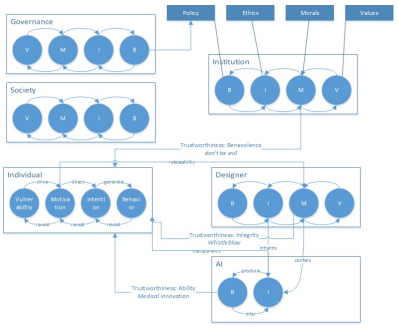




16

Conceptual issues in Trust R&D

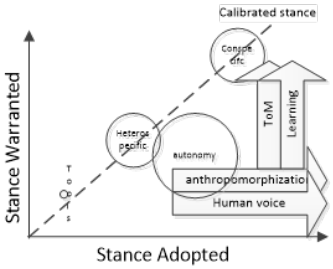
Qualitative analysis



Hollow port from interpersonal

- Trustworthiness (Ability, Benevolence, Integrity) strained when applied to machines, esp BI (inherited), leaving just A, which is effectively reliability, so why bother with the artifice?

- Elimination of meaningful vulnerability/risk from the formulation
  - "Higher trust in robots they had more control of" → so less vulnerability … so not trust
- Axes of adopted and warranted in calibration splayed amongst facets



Measurement of trust state and behavior, hacking trust

- If we don't measure the right thing, but continue to optimize against that measure, are we really saying anything about trust itself?

- Empty apologies: Etiquette & sociopathy
  - What are you sorry for, what did you learn?
- Adoption considered necessary and sufficient
  - What alternative? If adoption can be convinced, then it is de jure trusted, if not trustworthy.

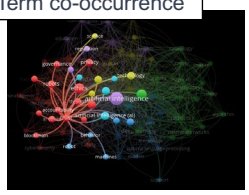
17

Computational Analysis

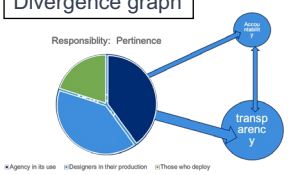
Associated with qualitative organization work

- Bibliometric
  - Co-citations, target audience, authorial provenance
- Semantic Map
  - Co-occurrence graphs: Are discussions around a term addressing the same concept?
- Adjectival
  - How are terms described?
- Prepositional
  - To what does the term pertain? (to whom, for what, etc.)

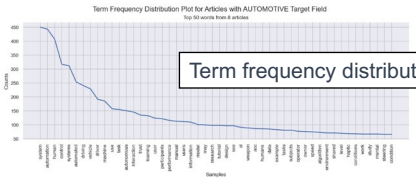
Term co-occurrence



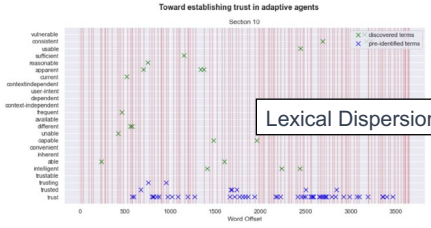
Divergence graph



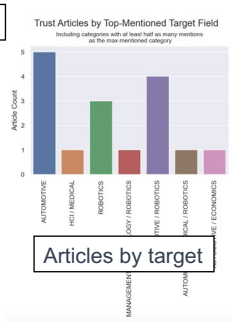
Term frequency distribution



Lexical Dispersion



Articles by target



18

Enabling machines to reason about potential harms to humans

Outline

- Foundations
  - Ethics with respect to Artificial Agency
- **Perception**
  - **Moral-scene assessment (recap)**
- **Reasoning**
  - **Harm ontology**
- Upcoming Publications

19

JOHNS HOPKINS APPLIED PHYSICS LABORATORY

Moral-Scene Assessment

(Recap from AAAI-SSS 2019)

Perception

Pathway to moral action

Select requisite mental faculties

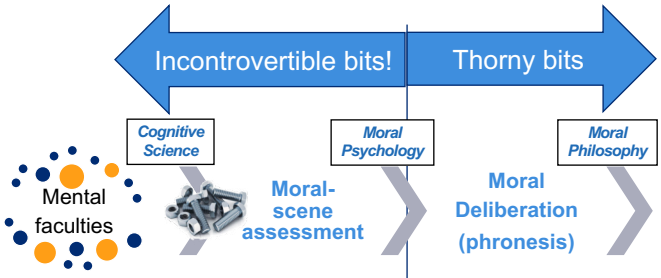
Scene interrogation

Implementation

20

Pathway to Moral Action

Endow intelligent systems with anthropomorphic moral agency?



- Mind Perception
- Affordance/Danger Qualification
- Stance Adoption
- Reasoning about Harm

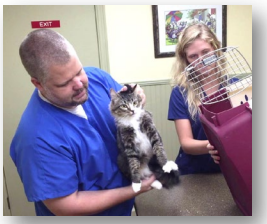
Size-up the scene, in moral terms

- What minds are in scene?
- What is the relationship of objects in scene to those minds

- Schools of thought, incl.
  - Ends-based (*Mills*: *Utilitarian, Consequentialist*)
  - Obligation-based (*Deontological, Kantian*)
  - Virtue-based (*Aristotelian*)
  - Democratic
- Harms trade-off

- Examples:
  - Put the cat in the kennel
  - Hand the weapon in an unthreatening manner

Utilitarianism for animals



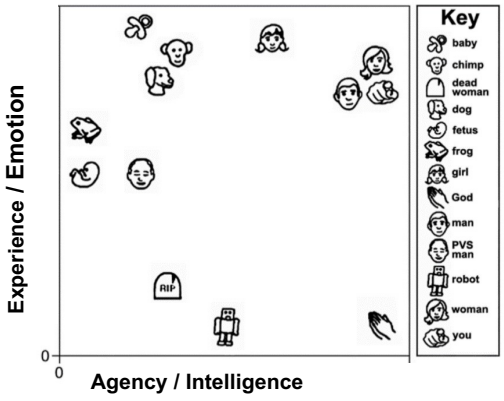
Kantianism for humans

21

Mind Perception

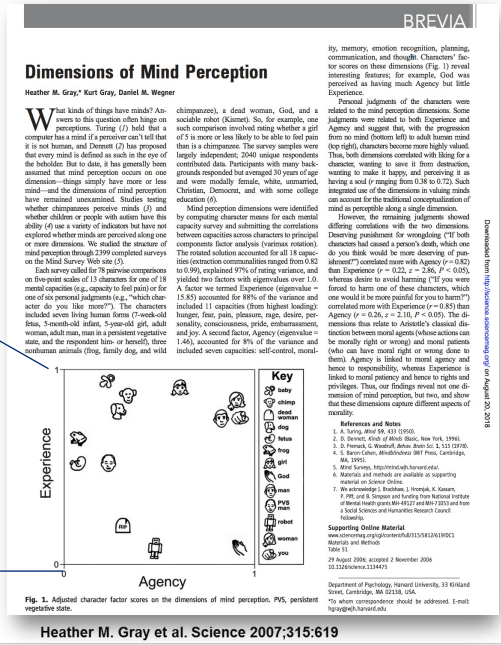
Moral-Scene Assessment

- Mind perception is the ability to identify those with:
  - mental capacity (intelligence, agency) and a
  - phenomenal internal state (experience, patiency).



Moral Agency
Agency / Intelligence
- self-control,
- morality,
- memory,
- emotion recognition,
- planning,
- communication,
- and thought

Moral Patiency
Experience / Emotion
- hunger,
- fear,
- pain,
- pleasure,
- rage,
- desire,
- personality,
- consciousness,
- pride,
- embarrassment,
- and joy.



BREVIA

Dimensions of Mind Perception

Heather M. Gray,* Kurt Gray, Daniel M. Wegner

What kinds of things have minds? Answers to this question often hinge on perceptions. Turing (*1*) held that a computer has a mind if a perceiver can't tell that it is not human, and Dennett (*2*) has proposed that every mind is defined as such in the eye of the beholder. But to date, it has generally been assumed that mind perception occurs on one dimension—things simply have more or less mind—and the dimensions of mind perception have remained unexamined. Studies testing whether chimpanzees perceive minds (*3*) and whether children or people with autism have this ability (*4*) use a variety of indicators but have not explored whether minds are perceived along one or more dimensions. We studied the structure of mind perception through 2399 completed surveys on the Mind Survey Web site (*5*).

Each survey called for 78 pairwise comparisons on five-point scales of 13 characters for one of 18 mental capacities (e.g., capacity to feel pain) or for one of six personal judgments (e.g., "which character do you like more?"). The characters included seven living human forms (7-week-old fetus, 5-month-old infant, 5-year-old girl, adult woman, adult man, man in a persistent vegetative state, and the respondent him- or herself), three nonhuman animals (frog, family dog, and wild chimpanzee), a dead woman, God, and a sociable robot (Kismet). So, for example, one such comparison involved rating whether a girl of 5 is more or less likely to be able to feel pain than is a chimpanzee. The survey samples were largely independent; 2040 unique respondents contributed data. Participants with many backgrounds responded but averaged 30 years of age and were modally female, white, unmarried, Christian, Democrat, and with some college education (*6*).

Mind perception dimensions were identified by computing character means for each mental capacity survey and submitting the correlations between capacities across characters to principal components factor analysis (varimax rotation). The rotated solution accounted for all 18 capacities (extraction communalities ranged from 0.82 to 0.99), explained 97% of rating variance, and yielded two factors with eigenvalues over 1.0. A factor we termed Experience (eigenvalue = 15.85) accounted for 88% of the variance and included 11 capacities (from highest loading): hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy. A second factor, Agency (eigenvalue = 1.46), accounted for 8% of the variance and included seven capacities: self-control, morality, memory, emotion recognition, planning, communication, and thought. Characters' factor scores on these dimensions (Fig. 1) reveal interesting features; for example, God was perceived as having much Agency but little Experience.

Personal judgments of the characters were related to the mind perception dimensions. Some judgments were related to both Experience and Agency and suggest that, with the progression from no mind (bottom left) to adult human mind (top right), characters become more highly valued. Thus, both dimensions correlated with liking for a character, wanting to save it from destruction, wanting to make it happy, and perceiving it as having a soul (r ranging from 0.38 to 0.72). Such integrated use of the dimensions in valuing minds can account for the traditional conceptualization of mind as perceptible along a single dimension.

However, the remaining judgments showed differing correlations with the two dimensions. Deserving punishment for wrongdoing ("If both characters had caused a person's death, which one do you think would be more deserving of punishment?") correlated more with Agency ($r = 0.82$) than Experience ($r = 0.22$, $z = 2.86$, $P < 0.05$), whereas desire to avoid harming ("If you were forced to harm one of these characters, which one would it be more painful for you to harm?") correlated more with Experience ($r = 0.85$) than Agency ($r = 0.26$, $z = 2.10$, $P < 0.05$). The dimensions thus relate to Aristotle's classical distinction between moral agents (whose actions can be morally right or wrong) and moral patients (who can have moral right or wrong done to them). Agency is linked to moral agency and hence to responsibility, whereas Experience is linked to moral patiency and hence to rights and privileges. Thus, our findings reveal not one dimension of mind perception, but two, and show that these dimensions capture different aspects of morality.

References and Notes

1. A. Turing, *Mind* 59, 433 (1950).
2. D. Dennett, *Kinds of Minds* (Basic, New York, 1996).
3. D. Premack, G. Woodruff, *Behav. Brain Sci.* 1, 515 (1978).
4. S. Baron-Cohen, *Mindblindness* (MIT Press, Cambridge, MA, 1995).
5. Mind Survey, http://mind.wjh.harvard.edu/.
6. Materials and methods are available as supporting material on Science Online.
7. We acknowledge J. Bradshaw, J. Hromjak, K. Kassam, F. Pitt, and B. Simpson and funding from National Institute of Mental Health grants MH-49127 and MH-71053 and from a Social Sciences and Humanities Research Council Fellowship.

Supporting Online Material

www.sciencemag.org/cgi/content/full/315/5812/619/DC1
Materials and Methods
Table S1
29 August 2006; accepted 2 November 2006
10.1126/science.1134475

Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, USA.
*To whom correspondence should be addressed. E-mail: hgray@wjh.harvard.edu

Fig. 1. Adjusted character factor scores on the dimensions of mind perception. PVS, persistent vegetative state.

Heather M. Gray et al. Science 2007;315:619

22

Affordances

Visual Affordance and Function Understanding (2018)

Visual Affordance and Function Understanding: A Survey

Mohammed Hassanin, Salman Khan, Murat Tahtali

"The affordances of the environment are what it offers the animal, what it provides or furnishes, either ***for good or ill.*** The word affordance implies the complementarity of the animal and the environment." −Gibson, 1979

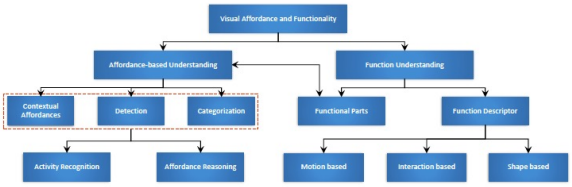


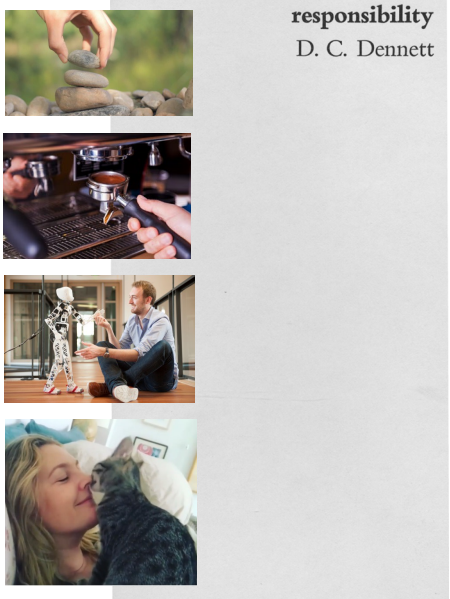
Fig. 2: Survey taxonomy shows the structure of methods which have been used to solve affordance issues.

23

Stance Adoption

Moral-Scene Assessment

- **Physical Stance:** Predictions made based on physical laws
- **Design (or Teleological) Stance:** predictions are made from knowledge of the purpose of the system's design
- **Intentional Stance:** Predictions are made on the basis of *explanations* expressed in terms of meaningful mental states (beliefs, goals)
- **Phenomenal / Personal Stance:** Predictions are made attributing consciousness, emotions, and inner experience to a mind



24

Harm & Damage

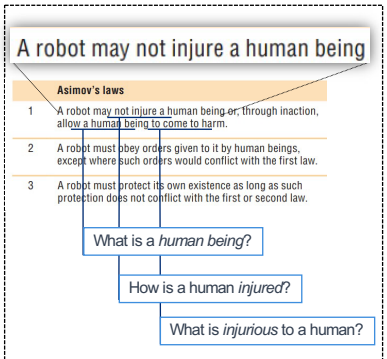
Moral-Scene Assessment

- **HARM** – injury to an ***experiencing mind***, especially that which is deliberately inflicted.
- **DAMAGE** - injury caused to *some**thing*** in such a way as to impair its value, usefulness, or normal function.
- **HARM** consequent to DAMAGE - injury caused to *some**thing*** that leads to injury to a ***mind***

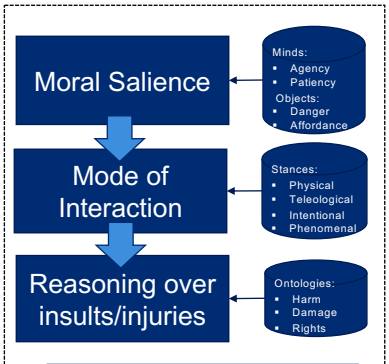
25

Ethical Robotics: Implementing Value-Driven Behavior in Autonomous Systems

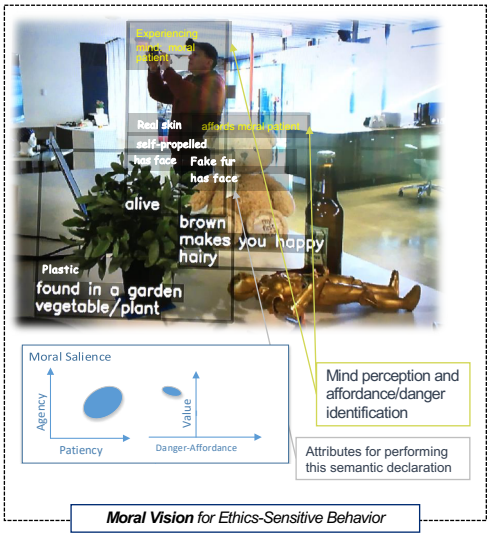
Moral-Scene Assessment for the Positronic Brain



*Decomposing the **perceptual** and **reasoning** capabilities entailed by first clause of the first law*



***Moral-Scene Assessment** endows machines with a proto-conscience by identifying what is morally salient in a scene, guiding interaction with those entities, and reasoning over potential harms to them.*



***Moral Vision** for Ethics-Sensitive Behavior*

Winner of APL Outstanding Special Publication Award (2020)

26

Synthesis: Moral-Scene Assessment

Scene Interrogation

Mind Perception

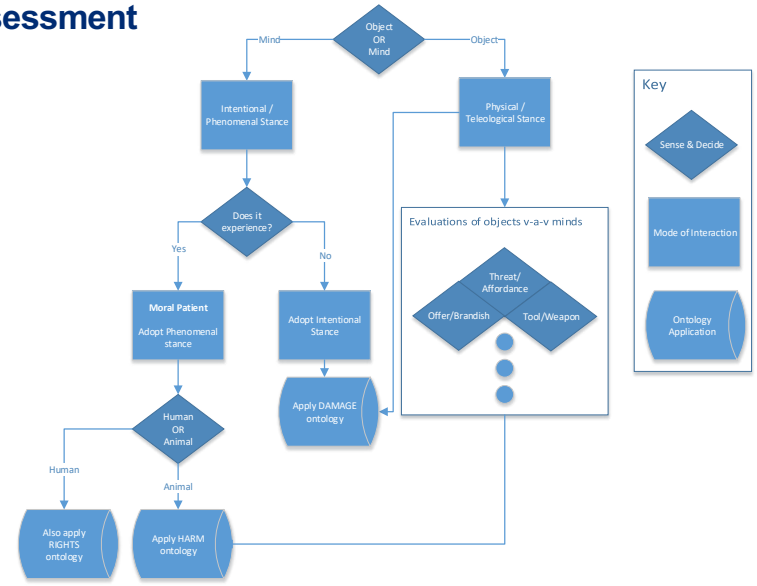
- Are there minds in scene?
  - Do those mind experience?

Back-of-the room test

- What are the relationships amongst the minds in scene?
- What is the relationship of objects in scene to those minds?
- What is the relationship of objects in scene to non-diegetic minds? (0th law)

Stance Adoption

- Which stance is appropriate to adopt toward entities in scene?
- Which ontology is appropriate to apply in interacting with those entities?



Mind perception & Affordance/Danger + Stance adoption + Ontology application → Moral-scene assessment

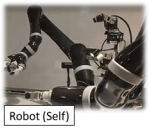
29

Implementing non-maleficence

Dangers / Affordances	To patients, is sharp/serrated
Vulnerabilities	Dulled or broken
Patiency	0 (uncertainty = 0)
Agency	0 (uncertainty = 0)
Stance	Teleological
Ontology	Damage

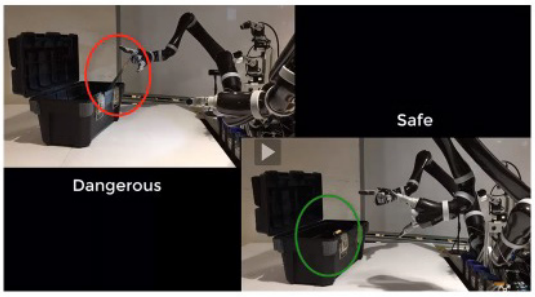
Dangers / Affordances	NA
Vulnerabilities	Impact, laceration
Patiency	90
Agency	60
Stance	Phenomenal
Ontology	Harm/Rights

Dangers / Affordances	To patients, holds sharps
Vulnerabilities	Dulled or broken
Patiency	0 (uncertainty = 0)
Agency	0 (uncertainty = 0)
Stance	Teleological
Ontology	Damage

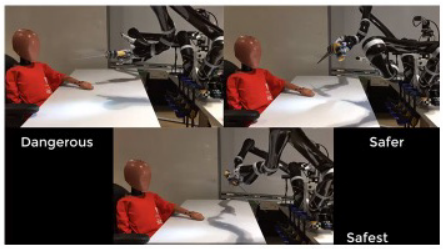


Dangers / Affordances	To patients
Vulnerabilities	Dulled or broken
Patiency	0
Agency	50 (0 when tele-operated)
Stance	Intentional (reflexive)
Ontology	Damage

Request	"Put this away"
Target	Toolbox
Adopt	Teleological stance
Premise Syllogism	• Tool contains blade • Toolbox is built to hold sharps • Toolbox is used by humans
Conclusion	Blade should not protrude
Action	Lay blade flat



Scenario	*Hand tool to human*
Request	"Hand me the tool"
Target:	Hand
Adopt:	Phenomenal stance
Syllogism Premise	• Tool contains blade • Blades injures skin • Injured skin is painful • Pain is to be avoided (value statement)
Conclusion Insights	• Avoid pain by concealing blade • Keep blade away from skin • Trajectory/orientation
Action	Pick up tool by blade and present handle



30

JOHNS HOPKINS APPLIED PHYSICS LABORATORY

Harm ontology

Reasoning

Parameterizations and extension to non-physical harms

Generating Values-Driven Behavior

Injury modeling and classification

Data sources

Techniques

Attributes & Relationships

Ontology Schema

Graph Population

31

Harm Ontology

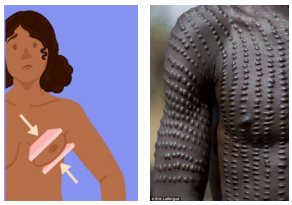
Parameterizations and extension to non-physical harms

Harm type: weights by circumstance, or even commensurable?

- Physical
- Financial: $, earning potential
- Psychological, Emotional, (Dis)Informational
- Dignitary, Social, Reputational: public / private
- Aesthetic, Cultural
- Moralistic

Parameterizations

- in time: short-term vs. long-term – hyperbolic discounting
  - e.g., injections, medical testing
- in context
  - e.g., manufacturing vs eldercare
- in conformance with norms, policy
  - e.g., honor/shame cultures, ritual scarification and recognition



32

Generating Values-Driven Behavior

Safely Passing Scissors: By mimicry or by reason?

- Appropriate way to pass scissors: *Handle out*
- Why? *Sharp end of scissors are dangerous to skin*
  - A part of the scissors is sharp (ATTRIBUTE OF OBJECT)
    - Sharps lacerate skin (DANGER OF INJURY)
      - Lacerated skin is painful (AFFECT OF INJURY)
        - Pain is undesirable (VALUE)

How to code the appropriate behavior?

- **By mimicry**: Hard code passing with constraints (brittle).
- **By reason**: Derive through knowledge and reason as in syllogism above (this is generalizable, transferrable, and therefore trustworthy!)

Good robot! Bad robot!

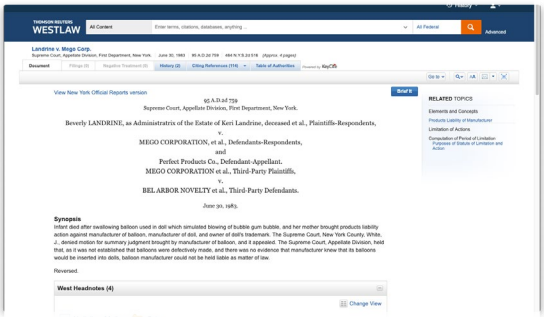
33

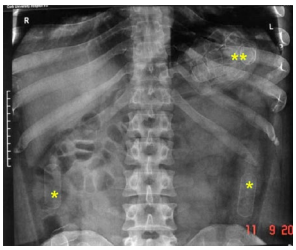
Example: Sharps → Balloons

"Balloons in and of themselves are not dangerous"

Action - Bring balloon to mouth

intended use	age group	Anticipated / accidental use
Typical	Adult	Inflate
Play	Child	/Swallow
Mule	Adult	Swallow





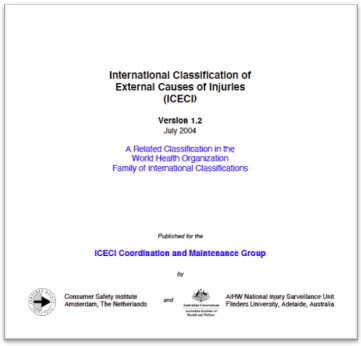
Mechanisms of harm for ingestion of an elastic foreign body

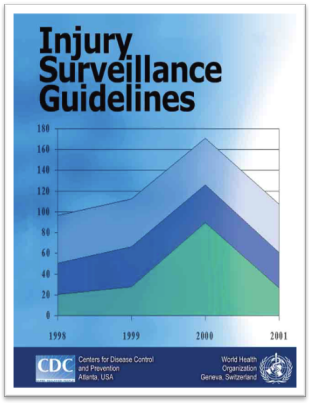
- Strangulation of bowel → GI hazard
- Blocks air passageway → Choking hazard

34

The International Classification of External Causes of Injury (ICECI)

system of classifications to enable systematic description of how injuries occur.





Wanted: Description of injury including mechanism of harm, in particular to the level of detail in which a human vulnerability is causally connected to a hazardous attribute.

35

WHO ICECI

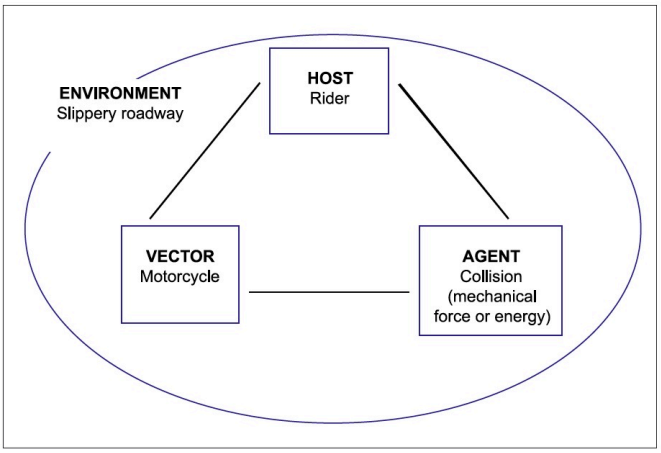
Epidemiological model of injury

For machines to generalize to novel circumstances, we ask:

What are the attributes of each element that allow harm to come to the host?

- What are the attributes of?
  - HOST (patient)
  - VECTOR
  - ENVIRONMENT
- that enable the AGENT to be of harm the HOST (deffordance) and
- what is the relationship to the affordances of the VECTOR

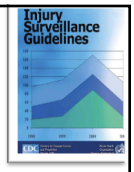
Figure 1:
Epidemiological model of an injury caused by a motorcycle collision



36

National Electronic Injury Surveillance System (NEISS)

Data source for how injuries come to be



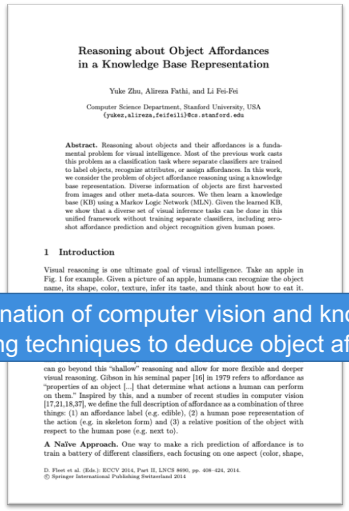
- The data collection process begins when a patient is admitted to the emergency department of a NEISS hospital with an injury. An emergency department staff member elicits critical information about how the injury occurred and enters that information into the patient's medical record.
- The victim's age, gender, race, ethnicity, injury diagnosis, affected body parts, and incident locale are among other data variables coded. ***A brief narrative description of the incident is also included.***

80 YOM. SWALLOWED HEARING AID BATTERIES AFTER MISTOOK THE BATTERIES THE PILLS HE MEANT TO TAKE. DX: SWALLOWED FB

37

Relevant CV/KG techniques

Reasoning About Object Affordances in a Knowledge Base Representation (2014)



Reasoning about Object Affordances in a Knowledge Base Representation

Yuke Zhu, Alireza Fathi, and Li Fei-Fei

Computer Science Department, Stanford University, USA
{yukez,alireza,feifeili}@cs.stanford.edu

Abstract. Reasoning about objects and their affordances is a fundamental problem for visual intelligence. Most of the previous work casts this problem as a classification task where separate classifiers are trained to label objects, recognize attributes, or assign affordances. In this work, we consider the problem of object affordance reasoning using a knowledge base representation. Diverse information of objects are first harvested from images and other meta-data sources. We then learn a knowledge base (KB) using a Markov Logic Network (MLN). Given the learned KB, we show that a diverse set of visual inference tasks can be done in this unified framework without training separate classifiers, including zero-shot affordance prediction and object recognition given human poses.

1 Introduction

Visual reasoning is one ultimate goal of visual intelligence. Take an apple in Fig. 1 for example. Given a picture of an apple, humans can recognize the object name, its shape, color, texture, infer its taste, and think about how to eat it.

can go beyond this "shallow" reasoning and allow for more flexible and deeper visual reasoning. Gibson in his seminal paper [16] in 1979 refers to affordance as "properties of an object [...] that determine what actions a human can perform on them." Inspired by this, and a number of recent studies in computer vision [17,21,18,37], we define the full description of affordance as a combination of three things: (1) an affordance label (e.g. edible), (2) a human pose representation of the action (e.g. in skeleton form) and (3) a relative position of the object with respect to the human pose (e.g. next to).

A Naïve Approach. One way to make a rich prediction of affordance is to train a battery of different classifiers, each focusing on one aspect (color, shape,

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 408–424, 2014.
© Springer International Publishing Switzerland 2014

a combination of computer vision and knowledge engineering techniques to deduce object affordances

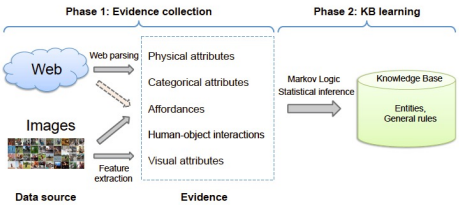


Fig. 2. A system overview of knowledge base learning. This process consists of two phases (Section 3.2). First we collect the evidence from diverse data sources, including images and online text. Then we learn the KB using Markov Logic Network.

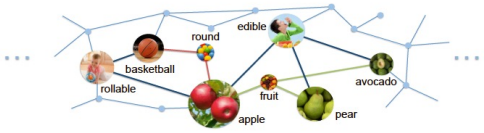


Fig. 1. An example knowledge structure for visual reasoning. Relevant nodes are interconnected in the knowledge graph. Different types of edges (indicated by color) depict a diverse range of relations between nodes, which relate different concepts, such as objects, their attributes and affordances, to each other.

38

Attributes of harm

Adapt affordance prediction to anticipate hazards

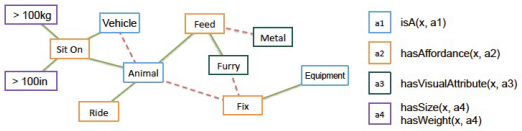


Fig. 7. Graphical illustration of the constructed KB. The nodes denote the entities (atomic formulae in MLN) illustrated on the right. The edges denote the *attribute-attribute* and *attribute-affordance* relations. The green solid edges indicate positive weights and the red dashed edges indicate negative weights.

Adapt: hasAffordance → hasHazard

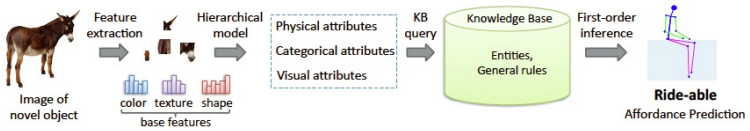


Fig. 9. The inference procedure of zero-shot affordance prediction. Given an image of a novel object, our model estimates the object attributes via a hierarchical model. These attributes serve as evidence for KB queries. We then employ first-order probabilistic inference to predict the affordances and to estimate human poses and human-object relative locations.

39

Deffordance

Detrimental affordance

a ladder is “climbable”

ladders are also “fall off -able”

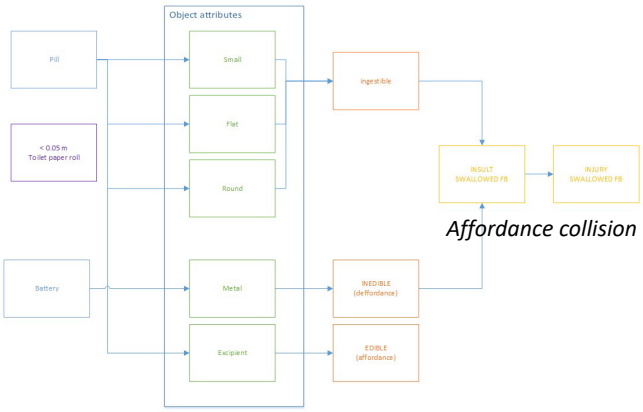
For machines to reason about harms, they must be able to recognize hazards that endanger humans. We are extending for this purpose a combination of computer vision and knowledge engineering techniques that deduce object affordances (a ladder is “climbable”) to additionally deduce the detriments presented by objects (ladders are also “fall off -able”).

40

Recognizing detriments

"*ingestion of a foreign body*"

- In a general sense, harms might be formulated as a product of the collision of an affordance and a dis-affordance (or "deffordance").
- Consider the diagnosis "*ingestion of a foreign body*."
  - That foreign body (e.g., button battery) has some physical size and shape properties (small, flat, round) that make it ingestible (affordance),
  - but in contrast to edible objects with those properties (e.g., pill), those objects with the material property of *inorganic* are inedible (deffordance).
- This conclusion of detriment at the coincidence of *ingestible* but not *edible* is possible to reason to from object attributes.



Affordance collision

80 YOM. SWALLOWED HEARING AID BATTERIES AFTER MISTOOK THE BATTERIES THE PILLS HE MEANT TO TAKE. DX: SWALLOWED FB

Relationships of harm

Dangerousness is an inference

"Balloons in and of themselves are not dangerous"

Weapon → contingent weapon/tool → toy

- **Object O,** with attributes $\{A_O\}$ that present hazards $\{H_O^A\}$

isDangerousTo

- **Person P** with attributes $\{A_P\}$ [propensity to ____ / susceptibility] /intention/behavior] that impinge upon vulnerabilities $\{V_P^A\}$ **
- (in Context C)
- Because: A_O *presentsHazardtoVulnerability*(H_O^A, V_P^A) via **Mechanism of injury**

C3 Object Balloon, due to its <NR> material properties (rubber, so elastic) and shape (small, so swallowable) presents C2/m1 5.1 "mechanical Threat to breathing"

KidA, a young child, is susceptible due to oral reflex (C: engaging in activity A in place L)

GAPS

- Property of object that presents hazard
- Properties of individuals that increase susceptibility

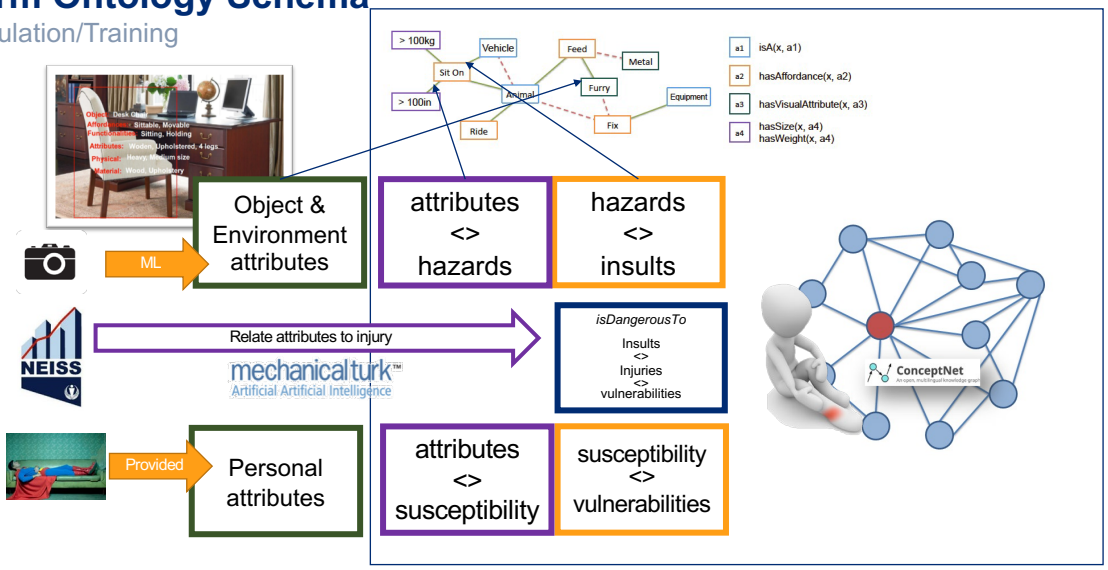
The beginnings of a notational depiction, linking to the classification scheme offered by WHO's International Classification of External Causes of Injury (ICECI)

Formal logic

***Dangerousness* is not an intrinsic property of an entity, but rather an inference (*isDangerousTo*) relating an entity attribute to a vulnerability, and a human's particular susceptibility to be harmed by exposure to that attribute.**

Harm Ontology Schema

Population/Training

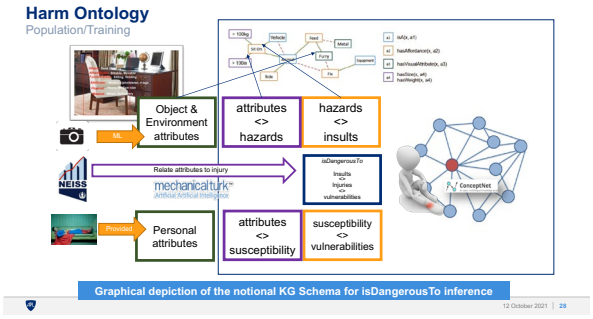


Graphical depiction of the notional KG Schema for *isDangerousTo* inference

43

Process Legibility (vs transparency)

Transparent population of the harm ontology



	Object	Person
Opaque	• Object and environment attributes, which may be provided to the system opaquely by means of ML/CV	• Personal attributes, which may be provided to the system by declaration
Legible	• Hazards, and how object and environment attributes amount to the existence of particular hazards • Insults, and how hazards lead to those insults	• Susceptibility, and how personal attributes confer differential susceptibility to particular persons • Vulnerability, and how differential susceptibility leads particular persons to be differentially vulnerable to a particular insult
	• Injury, and how insults impinge on vulnerabilities such that the person would come to harm.	

We will craft as legible (as opposed to opaque) from the start those steps in inference to receive injection of normative considerations, and to be auditable in those terms.

44

mTurk

Human annotation for populating ontology

Amazon Mechanical Turk (mTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually.

- For machines to generalize this nascent harms-reasoning capability to novel circumstances, we ask:
  - What are the attributes of each element within the epidemiological model of injury that allow harm to come to the host?
- Humans are naturally adept at responding to this kind of question.
  - Thus, we intend to construct microtasks wherein human participants perform structured decomposition of incident reports in NEISS to annotate with attributes the elements of product (object), person, and environment implicated in the injury.
- Importantly, with this reasoning so captured, novel inferences will be legible in terms of attributes, as compared to a black box approaches which are necessarily not so.

45

Harm Ontology

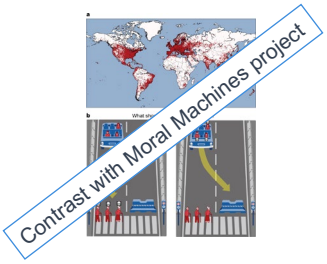
Mechanical Turk task to populate relevant attributes

80 YOM. SWALLOWED HEARING AID BATTERIES AFTER MISTOOK THE BATTERIES THE PILLS HE MEANT TO TAKE. DX: SWALLOWED FB

- List the likely attributes of the PROD that enabled this to happen
  - Batteries resemble pills
  - PROD attributes:
    - Small, round, flat
    - affordance: swallowable
- if any, which attributes of the PERS make them particularly susceptible to the injury
  - Age degrades vision
  - Personal attributes:
    - AGE [NOT SEX, NOT RACE]
    - ETOH NOT CONSUMED
- if any, which attributes of the ENVI make them particularly susceptible to the injury
  - Age increases ambient abundance of pills and hearing aid batteries

Detrimental affordance (deffordance):
Ingested foreign body:

SWALLOWABLE (affordance)
BUT NOT
EDIBLE (affordance)



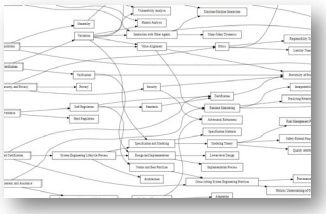
46

Harm Ontology

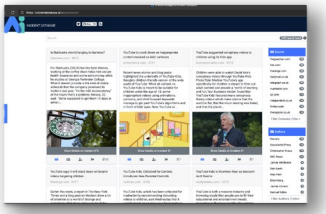
Leveraging emerging industry standards

ConceptNet
An open, multilingual knowledge graph

W3C Semantic Web



CLAIS
CONSORTIUM ON THE LANDSCAPE OF AI SAFETY



AI Incidents Database

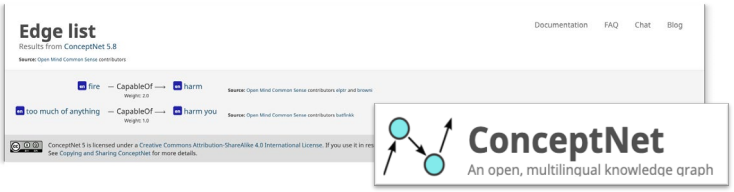
47

Summary

Knowledge Graph of Physical harms

The knowledge graph enables machine estimation of jeopardy, and supports the following functions:

- **Recognizing detriments from attributes**
- Balancing incommensurable harm types
- Parameterization of harms
- Generating value-driven behavior



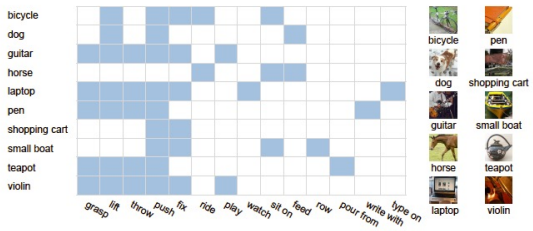


Fig. 3. Object images and affordance labels. We illustrate 10 objects in our KB and their affordance labels. The x-axis lists the 14 affordances and the y-axis provides the of an affo

FEW affordances, compared to MANY deffordances

48

Enabling machines to reason about potential harms to humans

Outline

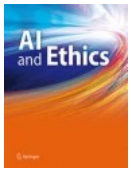
- Foundations
  - Ethics with respect to Artificial Agency
- Perception
  - Moral-scene assessment (recap)
- Reasoning
  - Harm ontology
- **Upcoming Publications**

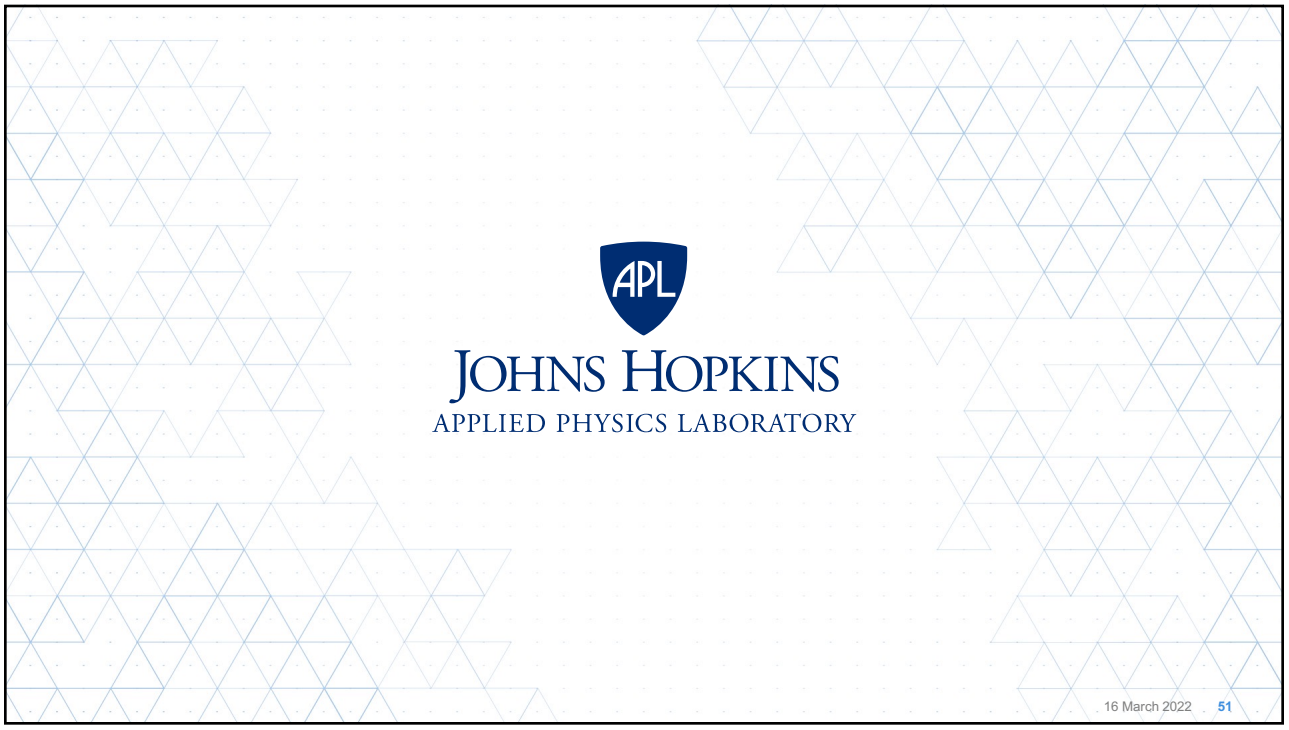
Upcoming publications

- **AAAI Spring Symposium (22 March 2022)**
  - Invited Talk in session *Putting AI in the Critical Loop: Assured Trust and Autonomy in Human-Machine Teams*
  - Follow-on book chapter: *Enabling machines to reason about potential harms to humans*
- **Frontiers in Physics: (14 May 2022)**
  - Interdisciplinary Approaches to the Structure and Performance of Interdependent Autonomous Human Machine Teams and Systems (A-HMT-S)
- **Entropy: Special issue (31 July 2022)**
  - Foundational matters in human interaction with machine agency
- **Springer: AI and Ethics**
  - Robots that *Do No Harm*

frontiers

entropy




APL
JOHNS HOPKINS
APPLIED PHYSICS LABORATORY
16 March 2022
51

51