Verification-Enhanced Learning for Safe Human-Robot Interaction

Shashank Pathak*, Luca Pulina** and **Armando Tacchella*****

(*) Istituto Italiano di Tecnologia (IIT)
(**) Università degli Studi di Sassari (POLCOMING - UNISS)
(***) Università degli Studi di Genova (DIBRIS - UNIGE)

14th AI*IA Conference, September 23-25, Ferrara, Italy

Robust AI: a long standing debate

'o appear, AAAI-94

The First Law of Robotics
(a call to arms)

Daniel Weld **Oren Etzioni***
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195
{weld, etzioni}@cs.washington.edu

Abstract

Even before the advent of Artificial Intelligence, science fiction writer Isaac Asimov recognized that an agent must place the protection of humans from harm at a higher priority than obeying human orders. Inspired by Asimov, we pose the following fundamental questions: (1) How should one formalize the rich, but informal, notion of "harm"? (2) How can an agent avoid performing harmful actions, and do so in a computationally tractable manner? (3) How should an agent resolve conflict between its goals and the need to avoid harm? (4) When should an agent prevent a human from harming herself? While we address some of these questions in technical detail, the primary goal of this paper is to focus attention on Asimov's concern: society will reject autonomous agents unless we have some credible means of making them safe!

The Three Laws of Robotics:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Isaac Asimov (Asimov 1942):

Motivation

In 1940, Isaac Asimov stated the First Law of Robotics, capturing an essential insight: an intelligent agent[1] should not slavishly obey human commands — its foremost goal should be to avoid harming humans. Consider the following scenarios:

- A construction robot is instructed to fill a pothole in the road. Although the robot repairs the cavity, it leaves the steam roller, chunks of tar, and an oil slick in the middle of a busy highway.
- A softbot (software robot) is instructed to reduce disk utilization below 90%. It succeeds, but inspection reveals that the agent deleted irreplaceable LaTeX files without backing them up to tape.

While less dramatic than Asimov's stories, the scenarios illustrate his point: not all ways of satisfying a human order are equally good; in fact, sometimes it is better not to satisfy the order at all. As we begin to deploy agents in environments where they can do some real damage, the time has come to revisit Asimov's Laws. This paper explores the following fundamental questions:

- **How should one formalize the notion of "harm"?** We define `dont-disturb` and `restore`— two domain-independent primitives that capture aspects of Asimov's rich but informal notion of harm within the classical planning framework.
- **How can an agent avoid performing harmful actions, and do so in a computationally tractable manner?** We leverage and extend the familiar mechanisms of planning with subgoal interactions (Tate 1977; Chapman 1987; McAllester & Rosenblitt 1991; Penberthy & Weld 1992) to detect potential harm in polynomial time. In addition, we explain how the agent can avoid harm using tactics such as *confrontation* and *evasion* (executing subplans to defuse the threat of harm).
- **How should an agent resolve conflict between its goals and the need to avoid harm?** We impose a strict hierarchy where `dont-disturb` constraints override planners goals, but `restore` constraints do not.
- **When should an agent prevent a human from harming herself?** At the end of the paper, we show how our framework could be extended to partially address this question.

*We thank Steve Hanks, Nick Kushmerick, Neal Lesh, Kevin Sullivan, and Mike Williamson for helpful discussions. This research was funded in part by the University of Washington Royalty Research Fund, by Office of Naval Research Grants 90-J-1904 and 92-J-1946, and by National Science Foundation Grants IRI-8957302, IRI-9211045, and IRI-9357772.

[1] Since the field of robotics now concerns itself primarily with kinematics, dynamics, path planning, and low level control issues, this paper might be better titled "The First Law of Agenthood." However, we keep the reference to "Robotics" as a historical tribute to Asimov.

The First Law of Robotics
[Asimov, 1940]
"A robot may not injure a human being, or, through inaction, allow a human being to come to harm."

"...before we release autonomous agents into real-world environments, we need some credible and computationally tractable means of making them obey Asimov's First Law."

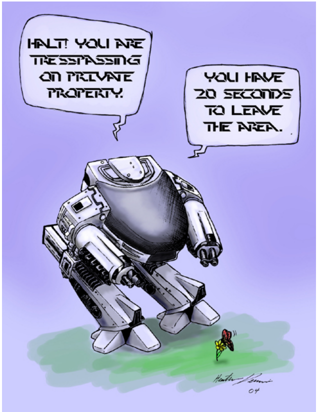
"Given a complex world where the agent does not have complete information, **any** *attempt to formalize the second half of Asimov's First Law is fraught with difficulties."*

Down-to-earth target: Dependable AI

Robots should meet **RAMSS requirements**:

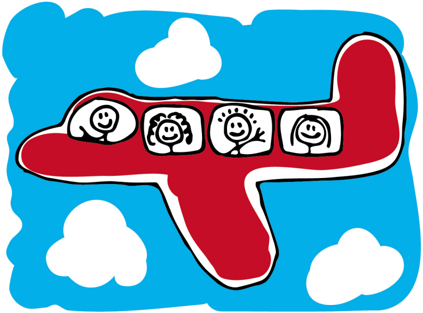
- **Reliability**: ability to perform required functions under stated conditions for a specified period of time
- **Availability**: proportion of time a system is in a functioning condition
- **Maintainability**: probability that a system will be retained in or restored to a specified condition within a given period of time
- **Safety**: ability to control recognized hazards to achieve acceptable level of risk
- **Security**: degree of resistance to, or protection from system damage

Autonomous robots are not, e.g., planes...



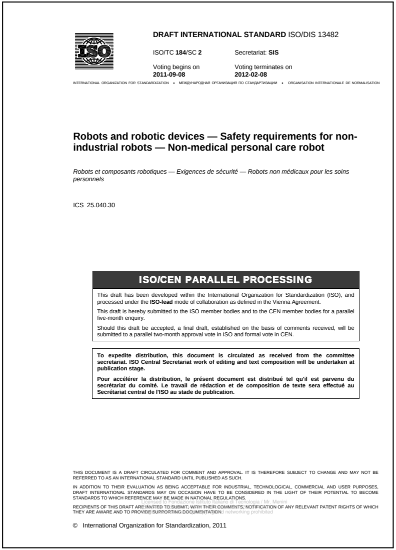
ED 209 shows a reliability defect, leading to potential safety defects

vs.



Planes are dependable, but we do not expect them to operate autonomously (if they did, they would be UAVs)

... still, they (will) need to be certified



DRAFT INTERNATIONAL STANDARD ISO/DIS 13482

ISO/TC **184**/SC **2** — Secretariat: **SIS**

Voting begins on **2011-09-08** — Voting terminates on **2012-02-08**

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION • МЕЖДУНАРОДНАЯ ОРГАНИЗАЦИЯ ПО СТАНДАРТИЗАЦИИ • ORGANISATION INTERNATIONALE DE NORMALISATION

Robots and robotic devices — Safety requirements for non-industrial robots — Non-medical personal care robot

Robots et composants robotiques — Exigences de sécurité — Robots non médicaux pour les soins personnels

ICS 25.040.30

ISO/CEN PARALLEL PROCESSING

This draft has been developed within the International Organization for Standardization (ISO), and processed under the **ISO-lead** mode of collaboration as defined in the Vienna Agreement.

This draft is hereby submitted to the ISO member bodies and to the CEN member bodies for a parallel five-month enquiry.

Should this draft be accepted, a final draft, established on the basis of comments received, will be submitted to a parallel two-month approval vote in ISO and formal vote in CEN.

To expedite distribution, this document is circulated as received from the committee secretariat. ISO Central Secretariat work of editing and text composition will be undertaken at publication stage.

Pour accélérer la distribution, le présent document est distribué tel qu'il est parvenu du secrétariat du comité. Le travail de rédaction et de composition de texte sera effectué au Secrétariat central de l'ISO au stade de publication.

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENT AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

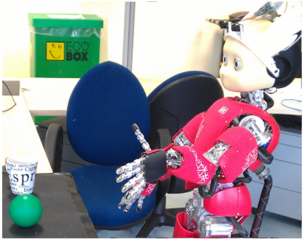
RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

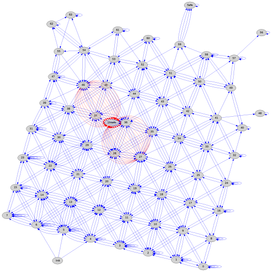
- ISO 13482
- Safety requirements for Non-industrial robots
- Non-medical personal care robots
- Makes provision for safe autonomous actions

Dependable AI <u>is</u> part of the picture

- **Intrinsic safety**: it is not possible to model an unsafe agent
  (Unlikely)
- **Safety by construction**: the agent will be safe as long as specific design guidelines are strictly observed
  (Staple method in engineering)
- **Demonstrable safety**: it can be proved that the agent design reduces undesirable events to an acceptable level
  (Our contribution!)
- **Monitorable safety**: it can be ensured that the agent recognizes actions leading to undesirable events
  (Hardly disposable, it comes next in our reserch agenda)

Verification and learning: competing needs!





Learning with discrete abstractions

Curse of dimensionality!

Learning in continuous state-space?

Verification on (discrete spaces

Can deal with large state spaces

Infinite state space makes verification undecidable!