

SAFE AND SECURE AI SYSTEMS

PRESENTATION FOR NVAIR 11 JUNE 2021

LEON KESTER

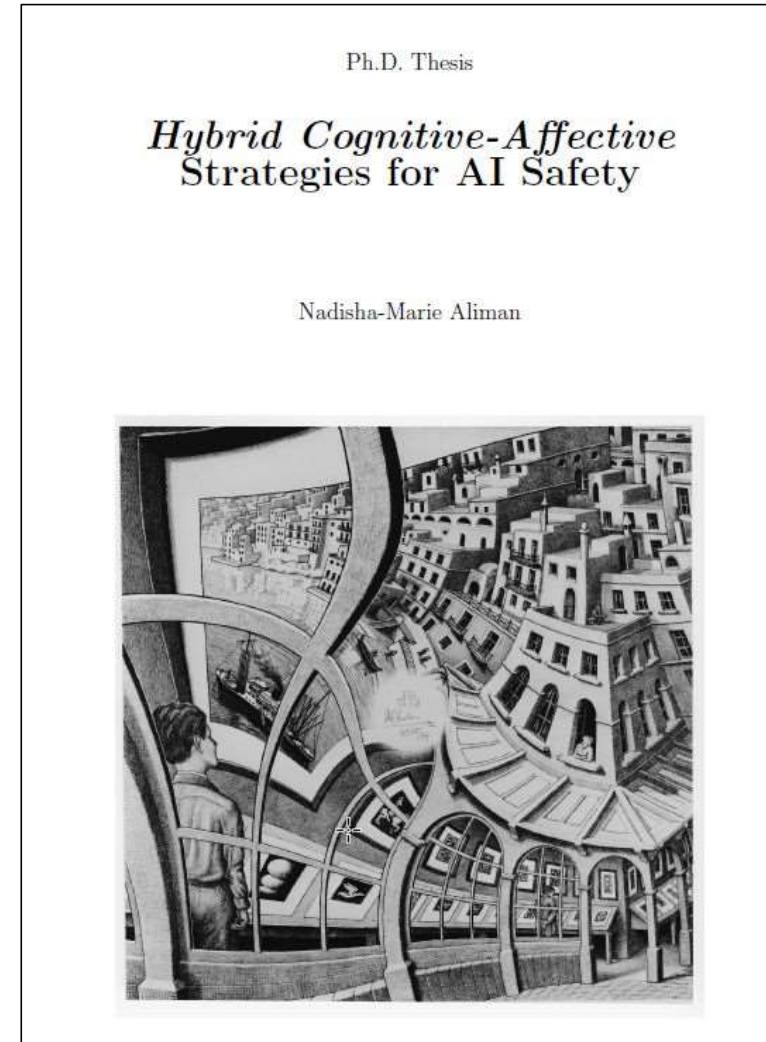
AI SAFETY

'COMMON UNDERSTANDING'

- › “It is instead first and foremost of relevance not to be afraid, not to be enthusiastic, but to **understand**” – Spinoza

- › Trans-disciplinary understanding of:
 - › AI Systems engineering
 - › AI and Cyber security
 - › Ethics, Law, Politics
 - › Philosophy of mind, Neuro-science, Psychology

- › Can AI systems create new explanatory knowledge?
 - › Philosophy of science



ARTIFICIAL INTELLIGENCE

› in well structured environments



1997



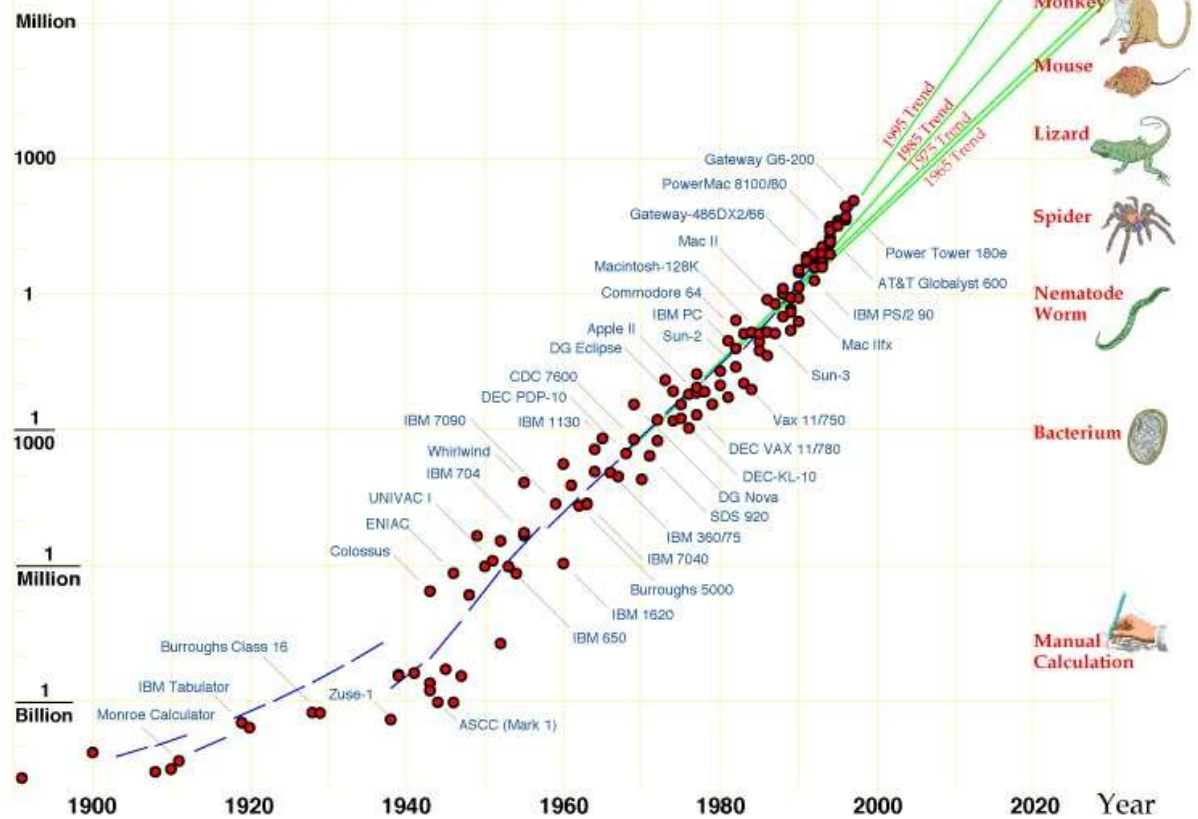
2016

COMPUTING POWER FOR ARTIFICIAL INTELLIGENCE



Evolution of Computer Power/Cost

MIPS per \$1000 (1997 Dollars)



ARTIFICIAL INTELLIGENCE

› in more unstructured environments considering “benevolent” humans.

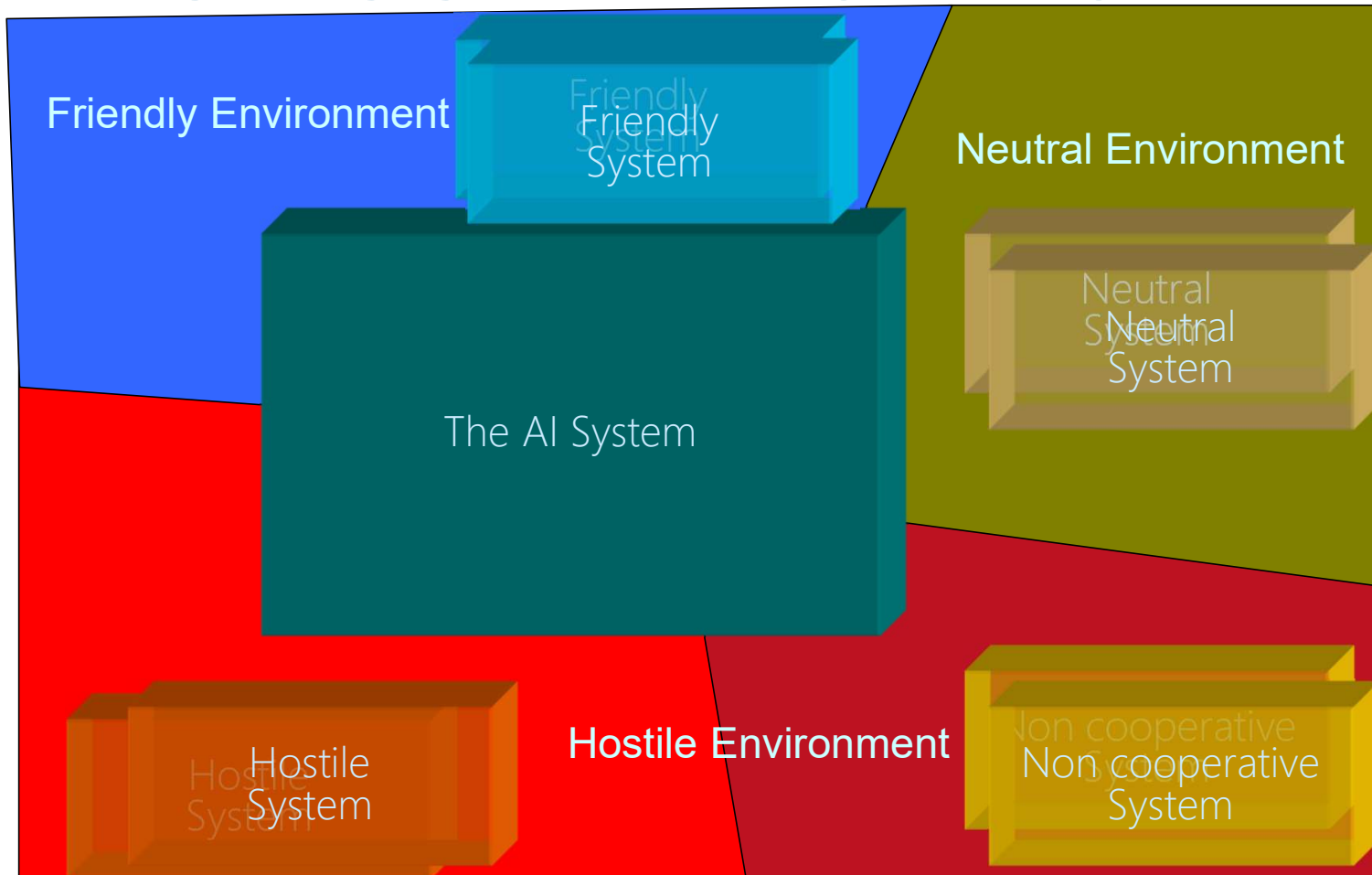


ARTIFICIAL INTELLIGENCE

- › Also considering “malevolent” humans.



INTELLIGENT SYSTEM IN AN OPEN WORLD



Kester L. (2010) Method for Designing Networking Adaptive Interactive Hybrid Systems. Interactive Collaborative Information Systems, vol 281. Springer, Berlin.

YAMPOLSKIY'S A(G)I SAFETY TAXONOMY

<i>How and When did AI become Dangerous</i>		<i>External Causes</i>			<i>Internal Causes</i>
		<i>On Purpose</i>	<i>By Mistake</i>	<i>Environment</i>	<i>Independently</i>
<i>Timing</i>	<i>Pre-Deployment</i>	<i>a</i>	<i>c</i>	<i>e</i>	<i>g</i>
	<i>Post-Deployment</i>	<i>b</i>	<i>d</i>	<i>f</i>	<i>h</i>

TWO TYPES OF AI LEADING TO AI SAFETY PARADOX

Error-Correction for AI Safety

Nadisha-Marie Aliman¹, Pieter Elands², Wolfgang Hürst¹, Leon Kester²,
Kristinn Thórisson⁴, Peter Werkhoven^{1,2}, Roman Yampolskiy³, and Soenke
Ziesche⁵

Type I AI: All current forms of AI

Type II AI: Hypothetical but scientifically possible type that can consciously create and understand explanatory knowledge (including ethics)

The AI Safety Paradox: *AI control and value alignment represent conjugate requirements in AI Safety.*

NEW PROPOSED A(G)I SAFETY TAXONOMY

TYPE I CLUSTER

<i>How and When did Type I system become Dangerous</i>		<i>External Causes</i>		
		<i>On Purpose</i>	<i>By Mistake</i>	<i>Environment</i>
<i>Timing</i>	<i>Pre-Deployment</i>	<i>a</i>	<i>c</i>	<i>e</i>
	<i>Post-Deployment</i>	<i>b</i>	<i>d</i>	<i>f</i>

TYPE II CLUSTER

<i>How and When did Type II system become Dangerous</i>		<i>External Causes</i>			<i>Internal Causes</i>	
		<i>On Purpose</i>	<i>By Mistake</i>	<i>Environment</i>	<i>On Purpose</i>	<i>By Mistake</i>
<i>Timing</i>	<i>Pre-Deployment</i>	<i>a</i>	<i>c</i>	<i>e</i>	<i>g</i>	<i>i</i>
	<i>Post-Deployment</i>	<i>b</i>	<i>d</i>	<i>f</i>	<i>h</i>	<i>j</i>

EP PROPOSAL FOR REGULATIONS FOR HIGH-RISK AI SYSTEMS - OBJECTIVES

- › “the Commission puts forward the proposed regulatory framework on Artificial Intelligence with the following specific objectives:
 - › ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values;
 - › ensure legal certainty to facilitate investment and innovation in AI;
 - › enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems;
 - › facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.

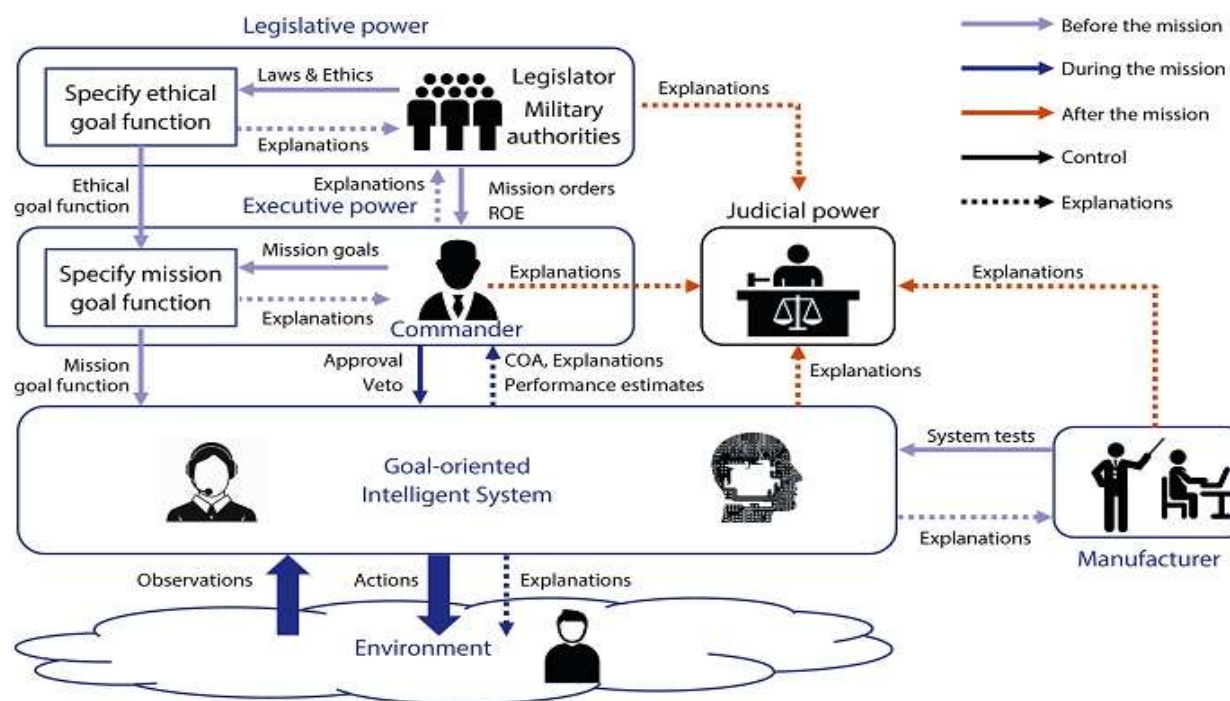
EP PROPOSAL FOR REGULATIONS FOR HIGH-RISK AI SYSTEMS – WAY TO GO

- › Option 3+: Horizontal EU legislative instrument following a proportionate risk-based approach + codes of conduct for non-high-risk AI systems;”
 - › A risk management system should be implemented

ISO definitions:

- › **Risk:** effect of uncertainty on objectives
- › **Safety Risk:** combination of the probability of occurrence of harm and the severity of that harm
- › **Harm:** injury or damage to the health of people, or damage to property or the environment -> Ethics

MEANINGFUL HUMAN CONTROL



Elands P, Huizing A, Kester L, Oggero S, Peeters M, Governing ethical and effective behaviour of intelligent systems, Military Spectator, 2019

PROBLEMS FOR IMPLEMENTATION OF ETHICAL AI

nature
machine intelligence

PERSPECTIVE

<https://doi.org/10.1038/s42256-019-0114-4>

Principles alone cannot guarantee ethical AI

Brent Mittelstadt ^{1,2}

Artificial intelligence (AI) ethics is now a global topic of discussion in academic and policy circles. At least 84 public-private initiatives have produced statements describing high-level principles, values and other tenets to guide the ethical development, deployment and governance of AI. According to recent meta-analyses, AI ethics has seemingly converged on a set of principles that closely resemble the four classic principles of medical ethics. Despite the initial credibility granted to a principled approach to AI ethics by the connection to principles in medical ethics, there are reasons to be concerned about its future impact on AI development and governance. Significant differences exist between medicine and AI development that suggest a principled approach for the latter may not enjoy success comparable to the former. Compared to medicine, AI development lacks (1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms. These differences suggest we should not yet celebrate consensus around high-level principles that hide deep political and normative disagreement.

THE AI CONTROL PARADOX

- A. In the EP proposal requirements for “Trustworthy” High risk AI systems are specified
- B. From a systems engineering perspective to be able to fulfill these requirements a **objective/goal** function specification is needed from which the AI system can derive the **risk** for every possible type of behavior and can do **risk** management.
- C. According to Ethicists and Lawyers an **objective/goal** function specification that includes **harm** in mathematical terms cannot be formulated satisfactory. Sometimes preference models or consequentialist utility functions are proposed but these functions encounter serious theoretical problems.

In order to fulfill A, either B or C (or both) must be false.

STEPS ON THE ROAD ...



Technology in Society

journal homepage: <http://www.elsevier.com/locate/techsoc>



Developing a roadmap for the moral programming of smart technology

Bart Wernaart

Fontys University of Applied Sciences, Eindhoven, the Netherlands

ARTICLE INFO

Keywords:

Moral programming
Ethical decision making
Moral intensity
Smart technology
Normative ethics
Meta ethics

ABSTRACT

Smart technology is increasingly integrated in our ethical decision making. This raises questions as to how we should morally program technology. Deciding on moral programming depends on the moral intensity of the ethical issue. A moral intensity dashboard for engineers can help allocate the most suitable moral authority for a particular moral programming. Technology is not capable of 'doing' ethics the way humans do. This leaves forms of consequentialism and deontology as the most reasonable programming alternatives, using deontic logic as a starting point. Furthermore, it is very likely that in the more complicated settings, technology should have elements of meta ethics in its moral programming to adequately deal with scenarios that lead to conflicts in moral programming. We propose to use the calculation methods that stem from a comparative approach or the Expected Moral Value approach. All this has considerable consequences in how we should see moral programming in technology-driven ethical decision-making processes. We will therefore propose a roadmap for the moral programming of smart technology.

MORAL PROGRAMMING

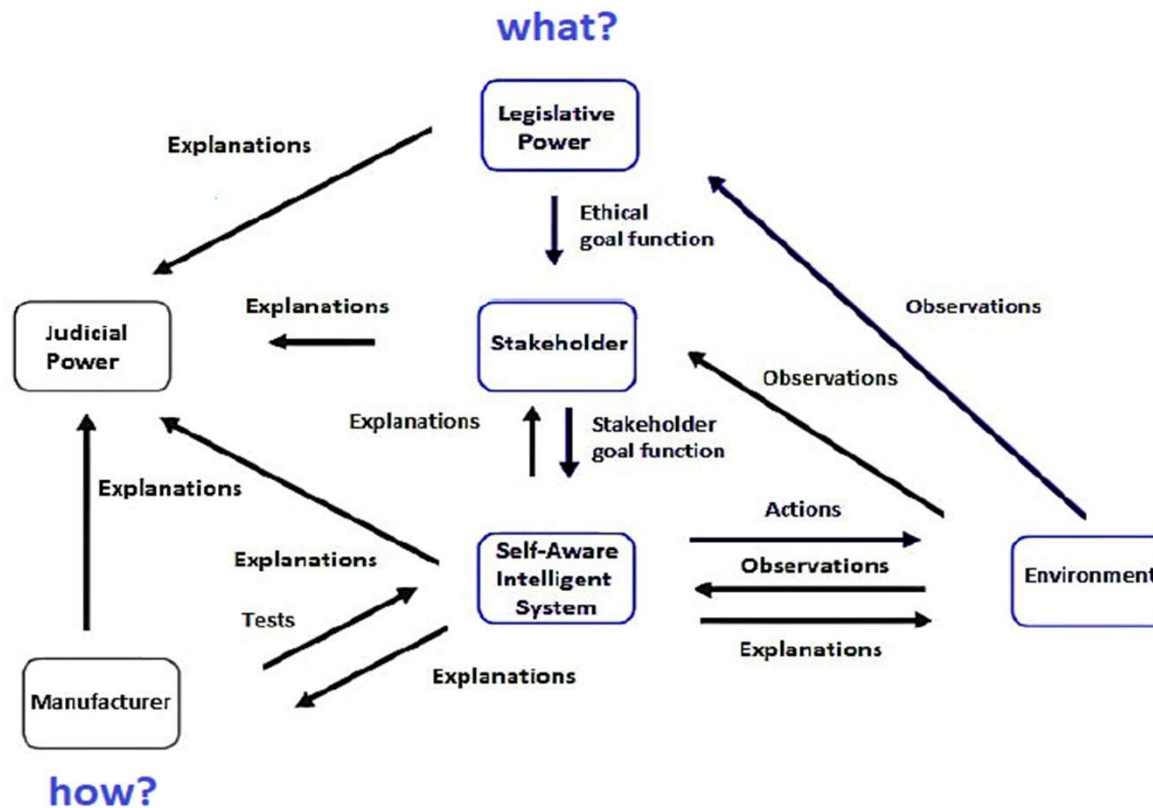
“The roadmap for moral programming” identifies promising approaches:

- Augmented Utilitarianism
 - "Augmented Utilitarianism for AGI Safety", Nadisha-Marie Aliman and Leon Kester, AGI-19, 2019

- Actor Deed Consequences model
 - Toward implementing the ADC model of moral judgment in autonomous vehicles, Veljko Dubljević, Science and Engineering Ethics 26 (5), 2461-2472, 2020

Chapter on Moral programming (Aliman & Kester) will appear in ***'Moral design and technology'***, Mr. Dr. Bart Wernaart (editor), Wageningen Academic Publishers

SOCIO-TECHNOLOGICAL FEEDBACK LOOP



Aliman, N.M., Kester, L., Werkhoven, P., Yampolskiy, R.: Orthogonality-Based Disentanglement of Responsibilities for Ethical Intelligent Systems. International Conference on Artificial General Intelligence, Springer (2019)

› **THE PRICE OF SECURITY IS ETERNAL CREATIVITY**

LEON.KESTER@TNO.NL