Systemic Deviation, aka: The Evil in The Machine

Paola Di Maio, PhD

ISTCS. ORG

paola.dimaio@gmail.com

Abstract

In the context of the increasing importance of transparency and accountability of intelligent and autonomous systems, this paper introduces identifies and defines the notion of Systemic Deviation and characterizes it with some examples, providing pointers to current and future work adopting this definition.

Introduction

Intelligent and autonomous systems are becoming easier and faster to develop and deploy, with increasingly lower barriers to achieve more advanced capabilities. Although systems automation has been the leading edge of industrial engineering for almost a century in all fields of application, systems autonomy capabilities are still evolving. The range and magnitude of implications for society are not yet fully understood. Responsible Autonomy and Robotic Ethics are two denominations for this broad field of study and practice. This paper starts by providing some background and context, defines Systemic Deviation providing some examples, proposes Joint Optimisation as a possible path to tackle Systemic Deviation, and points to further work being planned in this line of research.

Expanding Boundaries

Autonomous and semi autonomous systems have been in operation for decades across all sectors, from manufacturing to health, but what is possible to achieve autonomously by engineered systems is being constantly redefined and it can be assumed that not all advances are reported in the public domain, as not all research results are published. On planet earth, the leading edge of autonomous systems today can be found in driverless vehicles, robotic surgery, remote operations and possibly, self assembling systems, whereby machines or intelligent household (even furniture and toys) are equipped with built in capabilities of putting themselves together and can, to some extent, perform self diagnosis and self repair. Space operations are largely reliant on systems autonomy, as much of space equipment and vehicles are operated remotely. [1] As systems become more intelligent, autonomous and complex, the possibility of 'emergent behavior' which is by definition 'unpredictable', should be considered at design stage. Systems engineers generally do not make moral judgments, a system is 'good' if it does what it is designed to do correctly and efficiently. Yet a degree of responsibility, abiding to a code of ethics, is part of the profession. The majority of systems engineers especially when employed by defense, will not ask whether the systems they develop will be used according to politically correct deployments, or in 'just wars'. They develop systems designed to function correctly, and moral judgments could even eventually become an impairment to correct functioning. Not everyone agrees that it is the responsibility of engineers to build ethical systems. at the same time, there is a need for increased awareness as to what an engineers ethical responsibility may consist of. New capabilities demand for new responsibilities, take into account the context and boundaries of what is considered a system, and to tackled questions such as How can engineered systems be designed to function according to 'ethical' values to minimize the risk of unintended harm. Literature can help to provide an historical perspective on 'ethics', yet from a systems engineering point of view ethics is technically a set of rules, a catalogue of constraints generally implemented via logic and algorithms. In humans, it can be said that ethical considerations are motivated and triggered by 'higher cognition', that is, the awareness that humans are a responsible part of a whole, which includes other humans, other species, and the environment. Capturing requirements and designing' systems in consideration of widening boundaries, whether of other individuals, their cultures, of the environment, and building functionality that conforms to stakeholders stated intentions and actions is a critical part of the engineering process. In intelligent autonomous systems, such awareness needs to be modeled and crafted skillfully. The focus of this paper is to identify and a phenomenon defined here as ‘systemic deviation’.

Systemic Deviation

'Systemic deviation' can be defined as a behavioral and functional shift of a system from its intended purpose. It can occur when:

a) A system is designed with the intent to disguise its real aim to achieve a different goal from what it appears to be designed for. This is not 'versatility'. Using a car as shelter, as opposed to drive it, is not considered systemic deviation. For example software or hardware designed to spy or gather intelligence for military or otherwise hostile purposes but marketed as technology for human rights and peace mission to its developers and engineers, is an example of systemic deviation.

b) A system is deliberately designed to function in a different way rather than optimal. Deviation is different from 'malfunction' in that malfunction is never deliberate, while deviation can be intentional or not.

c) A system designed to function in a certain way for a certain (ethical) purpose, is misused and made to function to fulfill the exact opposite (unetical) purpose: for example a surgical scalpel, designed to perform medical procedures on patients to save their lives, vs the same surgical scalpel used to stab and murder a patient or any other person. This research at this stage is concerned primarily with a) and b), not with c) since the deliberate misuse of a system is not tackled at systems design stage in the system lifecycle. Systemic Deviation occurrences are widespread, the consequences visible in a variety of very common, easily observable examples:

- hospitals that kill patients by accident or mistake
- corruption in law enforcement agencies
- states and governments run by insiders who operate on behalf of private interests
- democracies which are actually governed and run by elites
- media disseminating disinformation
- seemingly progressive movements which are actually set up to stir and drive repression
- open data initiatives masterminded by secretive individuals and organisations behind closed curtains
- ships designed to sink (presumably by infiltrated enemies)
- transparency initiatives designed and led by people and organisations who want to make transparent models fail

From Human to Machine Deviation

Our age is characterized by technological progress, with engineered systems increasingly steering and controlling our lives, especially in terms of information and knowledge, and their influence on our behavior. In a philosophical context mental defilements, such as wrong assumptions, poor reasoning and incorrect conclusions leading to wrong decisions and to unintended outcomes, can be summarized as ignorance and wrong views and generally accepted as human fallibility.

‘Wrong’ views – generally resulting from incomplete information and knowledge – unless corrected. can be transposed and propagated from fallible human minds to pervasive engineered systems. Technology, especially critical technologies and precision systems have been developed according to rigorous methods and standards, but as barriers are lowered to generate widely accessible and distributed new capabilities, with simplified and automated processes quality assurance and system behavior controls are not necessarily upgraded accordingly.

Measures

A combination of approaches and diverse measures can attempt to tackle the complexity, in particular the sociotechnical complexities arising from these new challenges for engineers.

These may include:

a) refine and strengthen engineering design principles and philosophy, so that deviations can be either avoided or immediately recognized and addressed at concept definition stage and modeling, before they trickle down to system design and implementation.

b) refine and strengthen engineering methods and protocols to prevent accidental or intentional deviation from occurring, although misuse can never really be prevented.

c) make systems engineers and designers aware of the exponential risks derived from

overlooking the importance of accurate and detailed design and documentation, as well as the need of clarity and unambiguity of functional models and processes
d) Develop methods addressing underlying complexity

Devising and Implementing Systems Ethics

Some initiatives, such as Ethically Aligned Design for example [IEEE 2016] purport that some level of ethics can be designed into the system to prevent causing harm. Yet, none of these initiatives nor the resources they have produced so far actually serve this purpose. Probably due to the multidimensionality of the problem they aim to tackle, and the extent to which technical systems deviations largely depend on mental defilements and obscurations of the human mind. Paradoxically, EAD so far does not provide a mechanism to design ethically aligned systems at all, contrary to its name, and while critiques are welcome from the customer facing personnel of the standardization body, very little resonates in the form of addressing the shortfalls in the workgroups.

Human Behaviour

It is generally accepted that humans have inherent cognitive limitations and as such are fallible, and their nature is prone to 'deviation' – despite humans being highly intelligent and feature rich, capable of empathy and emotions, their ability to make best ethical decisions even in well meaning subjects can be dimmed by poor judgment, ignorance, second motives or simple human error. This may limit the ability of even educated and well intentioned individuals to develop design implement and use systems in such a way that these can cause malfunction and potential misuse.

Systemic Challenges

Developing methods for tackling systemic deviation encounters systemic challenges, intended as the systemic difficulties in tackling system dimensions.
The challenges to design ethical systems engineering include ensuring that human beings become capable of understanding (cogntiviely) and embracing (pragmatically) 'ethics', before they can become proficient at developing ethical systems. Yet even simply achieving consensus, or even just a shared working definition of what is ethics, is a major obstacle in the diverse and vocal communities attempting to do so. This is a fundamental problem that needs to be resolved

Socio Technical Joint Optimisation

Theoretically, to model Systems Ethics as a set of implementable behavioral rules for the purpose of reducing the risks posed by únethical' behavior, may be technically feasible. Yet, complex, highly dynamic socio technical systems require adaptation to gather and maintain optimal equilibrium. Socio Technical Joint Optimisation as defined in the Joint Optimisation Metamodel (JOM) [Di Maio, 2014] is achieved by adopting a design based on a set of core elements and extensions to support systemic joint optimisation. In its first version, JOM includes a provision to support different kinds of social norms, and could evolve to support System Ethics, by devising an additional extension to the metamodel. The concept of complex adaptive systems is crucial to an adequate understanding of the emerging field of complexity science. The concept represents the dynamic interactions of diverse agents who self-organize and produce adaptations that emerge in ways that can neither be predicted nor controlled

Research Agenda

Understanding and addressing systemic deviation is the latest chapter in the story of humanity. The work in hand includes
-codifying éxisting models of wrongdoing, as identified in the growing body of knowledge of whistleblowing research for example, and understanding how behavioural deviations become institutionalised, systematized, and embedded in technical systems. My work in this area includes developing a knowledge representation mechanism for systemic deviation and case based research and possibly methodological contributions for risk reduction
- Developing an Ethical Functional Model to support system development which adopts ethical values and norms as axioms

Conclusion

This paper introduces the notion of Systemic Deviation, and provides a working definition and some examples, and introduces a research agenda consisting of current and future work in this space

REFERENCES

Di Maio, P. (2014). Towards a metamodel to support the joint optimization of socio technical systems. *Systems*, *2*(3), 273-296.

Jónsson, Morris, Pedersen
Autonomy in Space Current Capabilities and Future Challenges
AI Magazine Volume 28 Number 4 (2007) (© AAAI)
https://pdfs.semanticscholar.org/56df/f5714a8d3b6f85903356ec91843af2ef50ae.pdf

IEEE https://ethicsinaction.ieee.org/