Theoretical framework for description and modeling of heterogeneous cognitive systems

This is the draft of the manuscript for discussion during HILL seminar. We kindly ask you to not distribute it further.
JZ, ŁJ, JRL

Abstract

We analyze processes of distributed cognition which in the digital age occur in multiple forms as interactions between humans and artificial agents. The starting point of our considerations is the enactivist approach to embodied cognition, which allow us to define agency in sufficiently general terms. Then we present a conceptual framework aimed at analyzing properties of heterogeneous cognitive systems consisting of multiple agents. It describes systems as sets of agents and their degrees of freedom bound by sets of constraints such as agents' physical capabilities, external conditions affecting their activity, communication protocols, shared cultural conventions, etc. Actions and communication of agents are modeled as an interplay of constraints and degrees of freedom, leading to the emergence of synergies and coordinations. The presented framework allows describing behavior of artificial and biological agents using common categories, while capturing important differences between their modes of operation. It affords a robust, analytical integration of some previous intuitions about possible models of interaction between human and artificial intelligence, and a framework for understanding composite, biological-artificial systems. We also propose how this enactivist general framework for understanding distributed cognitive systems gives rise to quantitative measures, useful for the empirical and computational sciences.

1 Introduction

For 25 years now we have been living as a part of a cognitive system called "the Internet". Initially, the technological part of it served us, humans, only as augmentation to our cognitive faculties of communication and content dissemination. With time, the digital mechanisms of the net shaped the way humans produced symbolic content and allowed for the explosion of new genres of representations. Presently, these mechanisms and techniques evolved into a new breed of artificial thinkers, mimicking (if not quite reproducing) some cognitive functions of human actors and introducing new ones. Our digital world cannot be thought of as "ours" alone anymore. We share it with various classes of artificial actors whose cognitive mechanisms and functions have been shaped by us, but which in turn continuously shape our thinking. Our understanding of the

world is, consequently, not ours alone; it is already now a product of symbiotic relationship of human and artificial cognitive agents and with the ubiquitous advent of the artificial intelligence (AI) it will become even more so.

This state of affairs inspires questions, among the general public as well as the scientific community, regarding the present and future relations between humans and artificial intelligence. Some people suggest that overreliance on digital technologies may lead to neglecting development of natural cognitive abilities that we already possess. Problems with attention span, memory etc. or symptoms of addiction may occur (Dworak, Schierl, Bruns, & Strüder, 2007; Kuss & Griffiths, 2017). Others are concerned about the possible harmful socio-economic impact of technological advancement, for example changes of the labour market due to automation ("Anticipating Artificial Intelligence," 2016; Campolo, Sanfilippo, Whittaker, & Crawford, 2018). This is connected with the fear that humans can be in some way replaced by machines, which will perform their task equally well, but without getting tired or making human errors. Finally, there are speculations regarding the possibility for artificial agents to reach or surpass human-level intelligence (Müller & Bostrom, 2016). Would such superintelligent agents share the same moral values as we do? Could they become hostile towards humanity?

We think that such questions are inherently difficult due to a lack of a conceptual framework that would be general and subtle enough to meaningfully address them. Even though it is generally agreed that on some level both human and artificial agents may be called "cognitive systems", when it comes to the specifics, they are described them differently. Artificial intelligence researchers and cognitive psychologists face different problems, use different vocabulary, and even when they borrow terms from one field to another those terms may have different connotations. Thus, it is not obvious how to study and talk about heterogeneous collective systems, consisting of diverse types of agents.

In this paper we aim to tackle the problem of cognitive heterogeneity. Our goal is to develop a theoretical approach that would allow to describe and model cognitive systems comprising coordinated agents of diverse cognitive origins and architectures (special case being human and artificial/digital agents, although our model should generalize to other composite systems: cooperating humans, humans and animals, or coupled artificial systems). Building a unified theory is not our ambition; presently we are satisfied with sketching out a meta-modeling framework guiding researchers constructing models of specific phenomena, and helping them to ask meaningful questions using a specific approach to cognition.

One could say that the information-processing framework for years did serve such a unifying goal. Treating human cognition in similar terms as computing machines, in principle, enables comparing them and describing interactions among them in a unified language. Yet, as many before us (Di Paolo, Buhrmann, & Barandiaran, 2017; Dreyfus, 1972, 1992; Searle, 1980; Varela, Rosch, & Thompson, 1991), we find this framework too limited to give justice to the richness of the "human side", ascribing to human cognition properties it might not have. We therefore seek a broader, less constraining, characterization of the cognitive systems, which does not start from treating them as certain intelligent artifacts. Unlike the information-processing framework, we do not assume any universal computational foundation to exist for all classes of cognitive processing (Turing, 1950).

Instead, we base on the enactivist and ecological psychology frameworks and

methods of dynamical systems to formulate an alternative framing for cognitive agents. This task consists of several important subtasks. In such a framework, in which the relation between the cognitive agents and their environments is constitutive for cognition, the independence of an agent from its environment needs to be questioned and the very identification of systems and their borders will become a relevant task. We will also need a characterization of cognitive agents that would abstract from recalcitrant philosophical issues and help a modeler to use it in concrete cases of human/AI cognition in real environments. Next, the general characterization of agents should naturally be suited to express interaction among agents and with the environment and provide a way to characterize composite agents, principles for aggregating agents and their scaling up in more complex systems.

Our framework will be congruent with radical embodiment perspective, advocating a full characterization of cognition through interactive processes and with the ecological psychology's insistence on agents' structured activity, being shaped by engagement in multiple projects. Dynamical systems theory, which is a natural choice for capturing multi-systemic and multi-scale interactive phenomena will lend its concepts for operationalization of the agent's functioning properties in terms of reduction of degrees of freedom under environmental and internal constraints.

We will develop our framework in the following steps:

Section 2. (Systemic approach to cognitive heterogeneity) provides a brief overview of the structure of our approach, our understanding of the notion of "system", its boundaries and modes of identification.

In Section 3. (Agents in the systems: communication, action, cooperation) we specify basic assumptions about the agents, ant their properties. We begin with a particular account of basic individual intentionality, then we explain the interplay between agents' degrees of freedom and various kinds of constraints, and then use these concepts to define system-wide mechanisms of communication, action and cooperation. The section ends with a handful of examples in which we use our framework to interpret functioning of various real-life cognitive systems, most of which are of heterogeneous composition.

Section 4. (Biological vs digital agents) shows the applicability of the systemic framework to characterize heterogenic cognitive systems. We start with expressing the cognitive specificity of these types of actors (stemming from differences in their "design processes") in our systemic terms of degrees of freedom and constraints. This allows us to analyse a number of various scenarios and compositions of heterogeneous human—AI assemblages.

In Section 5. (The future of artificial intelligence) we use our framework as a vantgage point to consider the main problems in human-AI interaction. Further, the same framework is applied to explain in more economical terms the postulates and goals of "Friendly AI" and various flavours of "Explainable AI", demonstrating the universality of our approach.

2 Systemic approach to cognitive heterogeneity

The framework designed for capturing cognitive heterogeneity should accommodate the variety of cognitive agencies and their compositions while at the same time preserving their specificity. The "common denominator" is needed to allow

formulating integrated accounts of cooperation between different kinds of "smart matter"—natural and artificial, human and non-human. The framework must be capable of including and operationalizing the important differences among agents as well as account for interactions among this variaty of agents in the descriptive accounts and models.

Our systemic approach is inspired by Bertalanffy's general system theory (von Bertalanffy, 1968). We define a system as an aggregate of agents acting on themselves and their environment. Although in most cases we will understand systems as consisting of various (e.g. human and non-human) classes of cognitive agents, it is important to underline that systemic approach allows for scaling and building hierarchies (and heterarchies accommodating more complex types of dependencies than just top-down), a key feature for understanding complexity of both physical and living systems (Pattee, 1972). Our "agents" themselves can be conceptualized as being complex systems of smaller components, and on the other hand, heterogeneous "system" itself can partake in meta-systemic structures. One of the goals of heterogeneous cognitive systems research should be an explanation of the rules and consequences of their scaling.

One of intended consequences of the systemic approach is avoiding essentialization of the differences between various classes of cognitive agents. These differences will be defined not by the variety of agents' "mental architectures", but will be expressed in terms of dynamics of a system as a whole. We want to be able to tell the differences between particular classes of agents by observing or modifying their relationship to the other components of the system. This way we will be able to avoid involvement in discussion about internal cognitive mechanisms and means of translation between different "mental architectures". The systemic dynamics and boundary conditions will serve as a common denominator and framework within which differences in cognitive properties and strategies will be observed. By doing this we depart from the internalistic paradigm of cognitive and mental architectures and from methods of artificial intelligence as means of modeling/reproducing human cognition, and adopt distributed and situated point of view which allows us to describe cooperation between the heterogeneous agents without defining the specifics of their inner workings.[1]

The minimal definition of a system above is treated as a general modeling abstraction applicable to a wide variety of phenomena. Borrowing from the field of agent-based modeling (Miller & Page, 2007), in our framework components of a system are "agents", again understood as convenient abstractions rather than beings equipped with agency in any strong sense. We characterize agent as autonomous (able to act independently), communicative (able to communicate with other agents) and active (able to perceive and act upon the environment). These terms reflect intuitions of the modeler, and may have different meanings depending on the modeled domain. Agents may represent individual people (sociological phenomena), larger aggregates like groups or independent countries (political science), or even communicating neurons or group of neurons (neurobiology).

[1] We wish to stress that we do not deny the existence of internal processes, their differences giving rise to the differences in (the behavior and experience of) cognitive agents. Our intention is to provide a framework which avoids the controversies involved in assuming certain mental architectures and therefore affords more practical and general descriptions.

System's boundary, environment

Agents acting in their environment constitute a system—a distinct whole consisting of parts. What constitutes system's wholeness? Or, in other words: how does a researcher building a model decides what to put inside the system, and what to leave out as part of system's surroundings? When thinking about physical and technical systems we often idealize them as closed systems. In such systems transfers of energy or mass in and out of the system are limited. The system itself may be then defined as a collection of certain physical bodies and their interactions. A sewing machine may be modeled as a system of mechanical components working within the machine case. With living beings and social systems, which are of particular interest to us, it is a different case. These kinds of systems should be conceptualized as open systems, in which there is a flow of matter and energy in and out of the system (von Bertalanffy, 1968). What remains constant are organizational properties of the system. Certain kinds of relations between agents inside the system and the outside world need to be maintained in order for the system to persist. System's boundary may be determined based on these relations.

Which organizational properties of a system one should consider necessary for system existence may be naturally a subject of deliberation when modeling a specific phenomenon. In the case of technical systems these properties will be connected with functions realized by the system. For example, a car engine needs to operate within a certain range of parameters in order to effectively move a car. External factors not affecting these parameters under normal circumstances may be left out of the modeled system (such as, for example, the presence of gravity, which we take for granted). In the case of self-organizing and self-regulating systems conditions necessary for system existence may be regulated internally. Autopoietic systems are defined as systems which actively construct themselves and consequently construct their own boundaries (Maturana & Varela, 1980). Autopoietic properties are characteristic for living organisms but also for many social systems (Zelený & Hufford, 1992).

To give an example of how a system's boundary might be determined let us consider what constitutes a sovereign state. Each state has some form of government, possesses authority over a certain geographical area, regulates life of its citizens, maintains diplomatic relations. In order to do so it needs to manage physical borders, register its citizens, introduce administrative divisions etc. Considering these regulations is possible to determine a natural boundary of the system.

3 Agents in the systems: communication, action, coorperation

In order to achieve neutrality in regards to agents' cognitive architectures we will describe them in terms of their interactions with environment (including other agents). Those interactions may happen across different modalities, but ultimately are constructed upon agents' basic dispositions to act. The repertoire of states out of which agent's interactions are composed—or, in other words, potential variability of agent's behaviour—will be encompassed by the notion of "degrees of freedom". On the other hand we will have "constraints", denoting

various ways of limiting agent's degrees of freedom or their expression, originating in the environment. Agents are differentiated by the dynamic interplay of degrees of freedom and constraints (Riley, Shockley, & Van Orden, 2012; Riley, Richardson, Shockley, & Ramenzoni, 2011).

The question that arises as the consequence of applying the above definition to our initial problem of heterogeneity of agents, is how do we distinguish different kinds of their agency in a way that will still allow for their compatibility as parts of a cognitive system. The special case we consider are systems consisting of human and artificial (digital) cognitive entities. Although there is no a priori reason why within the broader system definition and its problem domain the artificial agent could not be at least as autonomous, communicative and perceptive as its human counterpart, intuition suggests (and our initial problem requires) that there is actually a difference between these two types of entities. To conceptualize this intuition we turn to the notion of "intentionality", narrowing it down to a specific understanding as proposed by Merleau-Ponty; does intentionality—the existence or various modes or lack thereof—help us better account for the way heterogeneous cognitive systems work?

Merleau-Ponty's notion of intentionality

Traditionally, intentionality is considered a feature of a mind describing the ability of being "about something" – facts, entities, processes etc.; specific mental states (desires, beliefs, intentions), have the property of being "intentional". [2]

Here we use the notion of intentionality within the systemic framework congruently with the phenomenological approach of Merleau-Ponty (2002). He follows Husserl in distinguishing between reflexive "intentionality of act", and a more basic "operative intentionality", which is pre-reflexive and has to do with immediate involvement of the subject with the world without any intermediaries of mental states or consciousness. Intentionality is realized on the most basic level as motility (p. 158): concrete movements of embodied mind in space, directed towards the features of space. These movements encode "motor intentionality" or "motor project" which is the most basic, pre-reflexive way of relating to reality (p. 127).

The convenient aspect of Marleau-Ponty's point of view is that it provides us with a way of describing agents' intentionality without referring to their mental architectures but as function of their involvement with concrete, objective

[2] Searle (1983) allows for true, "original" intentionality to be attributed only to humans and ties it to their consciousness, expressed in their language of thought (Fodor, 1975). Other objects (sentences of natural language, machines, computer programs) can be intentional only by "borrowing" "aboutness" from the original source of (human) intentionality—their intentionality is derived. This account of intentionality would clearly distinguish types of actors in our case, however, it would also involve us in all the unresolved disputes about kinds of minds and the prospects of their communication given their varying cognitive architectures.

On the other hand Dennett (1989) claims that there is no reason to distinguish original and derived intentionality given that the former itself is a result of evolutionary design process. He advocates assuming "intentional stance": using intentionality as a practical framework for explaining and predicting behaviors when there is not enough knowledge about the way the agent had been designed or what the mechanism that govern its work on a material level are. The observed agent/system does not need to have any kind of consciousness, still, there is a meaning associated with its actions—contained in mental states implied by intentional stance. Application of this strategy to our problem of agents' heterogeneity is straightforward: intentional stance allows for treating equally all kinds of cognitive agents. Human intentionality is essentially no better or worse than that of an animal or computer.

outside world. The relation of an actor with its environment then becomes a common denominator; it allows us to account for heterogeneity of actors using common set of notions which operationalize the directedness of subject towards elements of the world.

The notion of bodily motility as an example of basic "operative intentionality", proposed by Marleau-Ponty, can be operationalized in a way we can use in descriptions of our rudimentary cognitive systems. Let's consider a simple agent placed in the objective world defined by one dimension of space (i.e., agent is placed on a "line") and one traditional dimension of time. The operative intentionality of agent in this situation will be defined as a set of "choices" it can "make": either to move "left" or "right", with "chosen" speed of movement. The increase of the complexity of both the agent and the environment results in an increase of "choices" which describe the potentiality of interactions. According to phenomenological tradition, these choices and interactions explain the very constitution of subject, all the way to reflexivity and consciousness. As Merleau-Ponty puts it, quoting Husserl: "consciousness is in the first place not a matter of 'I think' but of 'I can'" (p. 159).

Intentionality = Degrees of Freedom

We will conceptualize the above theoretical propositions—the range of agent's choices, modes of its directedness towards things other than itself (intentionality in Merleau-Ponty's sense), reality other than its own agency—as specific organizations of agent's degrees of freedom. In dynamical systems degrees of freedom correspond to independent parameters required to describe dynamics of a system. Systems with a larger number of degrees of freedom are considered more complex and may display more variability in their behaviour, as space of states which are possible for them to visit has more dimensions. Depending on the nature of the described system degrees of freedom may refer to different properties. For example, in kinematics degrees of freedom of a system consisting of a single molecule moving through space will encompass its position and velocity. In statistics degrees of freedom refer to possible ways a model may disagree with the data. In the context of communication, degrees of freedom might denote, for instance, the number of symbols that are available in vocabulary; in the case of action, the number of courses of action that can be taken by agents. Degrees of freedom and their organization will be useful to characterize the whole distributed system as well as individual agents.

The two kinds of actions

We distinguish particular functional organizations of agent's degrees of freedom as agent's actions. Let us recognize that some of the agent's actions affect primarily the system itself (what is inside the system boundary), most notably other agents, while other actions affect primarily system environment (what is outside the system boundary). The first kind of actions will comprise internal activity of an agent, while the second kind will comprise external activity of an agent. Later, this distinction will allow us to characterize different agent roles within a system, for example "managers", focused on the internal activity of regulating the system, and "field workers", performing tasks in the environment.

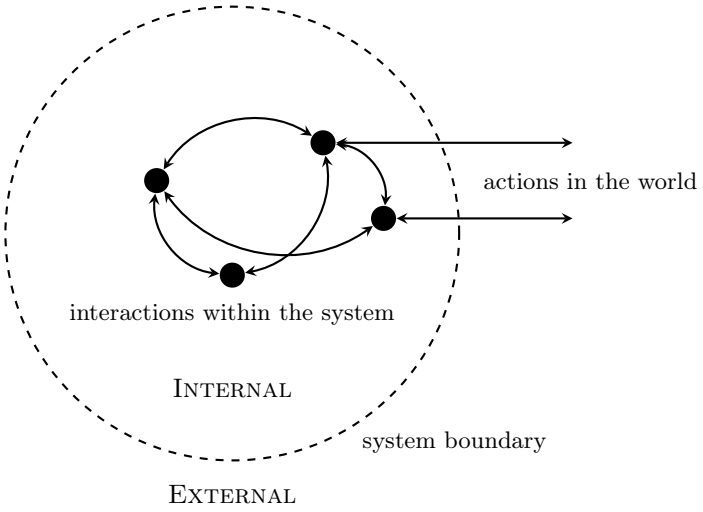


Figure 1: Agents acting within a system. Agents are represented as dots, their activity is divided into internal and external based on the system boundary.

The notion of constraint

Agent's degrees of freedom (and thus its activity) are constrained by its environment, by other agents, but also by its history enfolding on different timescales: individual history (accumulated knowledge, experiences), cultural history (customs, language), or biological evolution history (resulting in certain form of a body and physical capabilities). Some of these constraints are shared between agents, which affects the way their individual degrees of freedom may be coordinated when they form a collective system. For example, agents which share cultural constraints in the form of a common language have tools for a more precise coordination of their actions, which may lead to better division of labour. Agents which the same design of perceptual apparatus will be sensitive to the same environmental stimuli and thus will be able act as a backup for each other in some situations etc.

We will introduce two functional categories of constraints depending on which kind of activity they primarily influence. Internal constraints will be those regulating internal activity, while external constraints will be regulating external activity (naturally, these sets are not fully disjoint as some constraints may greatly influence both kinds of actions). Internal constraints facilitate system self-regulation and coordination between agents, often through providing shared communication protocols that the agents use. For example, introducing communicative devices (e.g. walkie-talkie) to a search party dispatched in a forest would be modeled as a modification of their set of internal constraints. External constraints affect agents' interactions with the environment; they dictate which actions are possible in a given situation (often due to physical capabilities of an agent). Providing maps of the terrain to the above-mentioned search party would modify their external constraints, allowing them to navigate the

environment more effectively. [3]

Cooperation, communication, action

Characterisation of agents as autonomous, communicative and active, introduction of degrees of freedom and various constraints, allows us to operationalize the notion of cooperation. It can be defined as coordinated action, resulting in pooling agents' resources (formalized as the ability to bind others' degrees of freedom), oriented towards realisation of a common goal. In other words, cooperation requires coordination of agents within the system (internal activity), which results in particular external activity congruent with the goal.

While distinguishing external and internal activity and constraints is useful for analysing system dynamics it is important to stress that in our framework the distinction between communication of agents and their actions is fuzzy. Defining communication as performative, influencing agents and environment, constructing—rather than just transmitting—meanings, allows for more effective conceptualising of cognitive differences between various classes of actors.

In our conceptualization of a system, the result of both communicating and acting can be described in terms of reduction of the number of degrees of freedom which characterise agents. Communication and action provide additional information for the agents, change the context of their involvement in the system and in consequence construct constraints and influence the space of choices available to them (Fusaroli, Rączaszek-Leonardi, & Tylén, 2014) (either restricting choice or opening up new possibilities). From a certain point of view acting is communicating, as it creates meanings by influencing behaviour of other actors—and vice versa, every communication among agents requires an action (a concrete act of "releasing" constraints into the world).

Ultimately, all kinds of constraints—internal and external, originating from communication or action, constitute a framework and resources for system's cognitive functioning. As agents communicate with each other, use environmental resources, and coordinate their action, they continue to mutually bind their degrees of freedom, forming functional synergies, and thus further refine system's cognitive capabilities towards the achievement of its projects and goals. Agents serve as a source of agency that shapes cognitive scaffolding, create a systemic niche within which thinking about solution of a given problem grows more plausible and efficient.

Examples

To better understand basic notions of our conceptual framework it is worth to apply them to some simple toy problems. The framework is inherently meta-

[3] It is interesting to note that internal and external constraints tend to be of different character. Internal constraints connected with the activity directed at other agents are often based on social conventions which evolve on the cultural timescale. They may have abstract and symbolic form. Such constraints are relatively fluid and it is easy to imagine that they will be modified and adapted during the repeated interactions of the same group of agents constructing their own social niche. External constraints, on the other hand, represent limitations of interaction with environment, which usually are more rigid by nature. They affect interaction with the environment and external objects. For instance, agent will require certain strength to move a heavy piece of rock. Such constraints may be also subject to change, for example due to biological evolution or through the invention of new tools and devices, but these processes occur generally on slower timescales.

theoretical, which means that it often describes modeling choices rather than direct properties of the modeled phenomena. As we will see, the same phenomenon might be often approached differently.

Construction workers A group of people working at the construction site constitutes a system. Each worker is an individual agent. They share a common goal of building a house. They share some constraints in the form natural spoken language as used in everyday situations. This, however, proves inadequate: in the noisy environment of the construction site they have difficulties hearing each other. Because of this, they gradually adopt a specialized repertoire of gestures facilitating communication, which are incorporated into their internal set of constraints. These gestures are grounded in the particular experiences of this group and their goal of building a house.

Shepherd and shepherd's dog Collaboration between humans and animals can also be described using this framework. A shepherd relies on the help of a dog to herd the flock of sheep. Shepherd and his dog are the agents in the modeled system (this is a modeling choice), the flock is part of their environment (another modeling choice is to ignore the agency of sheep and treat the behavior of the flock as an external process that needs to be regulated). Both the shepherd and his dog move in space in relation to each other but also in relation to the flock, shepherd shouts commands to the dog, the dog barks at the sheep etc.—potential variability of these basic actions is captured by agents' degrees of freedom. Communication between the shepherd and the dog, both in terms of verbal commands and in terms of positioning themselves in space, constitutes the internal activity, the actual herding is the external activity. There are internal constraints which facilitate communication—the dog understands from experience simple commands given by the shepherd, the shepherd reacts to dog's actions. The task of herding sheep imposes external constraints on their behaviour: for example, they cannot move too fast or they would lose some sheep. It is possible to measure the complexity (related to the number of degrees of freedom after applying all constraints) of such system. For instance, we could look for orderliness of the trajectory of the centre of a dog-human system in the task of herding and compare it with a situation when they just go for a walk.

Engineer and his tools Is an engineer using a set of different tools in the same situation as the shepherd? We would probably be hesitant to talk about a collaboration between human and hammer: a hammer is a simple tool fully controlled by its user. Possession of a hammer can be modeled as simple extension of an action repertoire of its user (enabling constraint); in most cases there is no need to attribute agency to a tool when it is used in a simple, unequivocal context. On the other hand, if we were to consider a different time-scale and ask how this hammer was designed, we would find ourselves analysing a much broader system consisting of an engineer and people designing hammers, who meet across time and space through the material object. The hammer would be in this case a carrier of agency of its designers. We could replace the tool in question which a sophisticated piece of CAD software, reacting to its user and suggesting its own solutions. This tool has more degrees of freedom and its interaction with the user (the way they mutually bind their degrees of freedom)

is much more complex and unpredictable. Because of this we will be more inclined to model it as a form of collaboration between human and artificial agent. Ultimately, the decision what to model as another constraint and what to model as an autonomous agent is left to the modeler, but it is based on the number of degrees of freedom of components and the complexity of their interactions.

Neural network Let us imagine an artificial neural network which is used as a control mechanism in an autonomous robot. The network has an architecture of multilayer perceptron (MLP) with one hidden layer, the input layer connected to robot's sensors, and the output layer connected to robot's actuators. Within our framework we will model each artificial neuron as an individual agent. Connections between neurons (network topology) may be described as internal constraints of this system. The immediate external environment of the network includes sensors and actuators connected to the input and output neurons, but in the broader context factors such as physical properties of robot's body, its environment, its purpose, etc. need to be included. Neurons from the input and output layers participate in both internal and external activity, while neurons from the hidden layer participate only in the internal activity. This emphasizes the key fact: even though all neurons are essentially identical and use the same activation function, they perform different roles due to how their degrees of freedom are constrained. The system's function is described mostly through its internal activity. If we were to model behavior of a swarm of such robots we would probably designate a single robot, not a single neuron, a unit of analysis—an agent. This illustrates how our framework may be applied to multiple scales of description of the same phenomenon.

Social network AI In social media content presented to the user is highly customized. We can portray the user and the algorithm generating her news feed as a collective system of two agents. The external activity of the user is related to her interests and manifests as her network activity: pages she visits, comments she writes, content she shares, etc. The algorithm operations are subject to external constraints representing goals determined by the social portal owners and its capabilities of content filtering. There are internal constraints facilitating communication between the user and the algorithm in the form of customization settings available to the user (for example, user may be allowed to limit the amount of personalized advertisement). However, all the "actions" of the agents are also meaningful in terms of communication: user's activity influences the algorithm, and algorithmically chosen content influences future behavior of the user. This means that in this case internal and external activity overlap greatly. This toy model could be expanded further by incorporating other users of the network, their personalized content filters, and their interactions.

Game of chess In chess two players engage in an intellectual struggle on a chessboard. Even though agents compete against each other (they have different goals within the game), their behavior is very much coordinated (they have a common goal of playing the game), thus they constitute a collective system. Their action repertoire consists of legal moves on the board. All actions of one player are directed at the other player, and their effects are limited to the board, the external world is not affected. Players react to each other moves, so

moving pieces on the board may be also interpreted as a form of communication between them. Chess is an example of a game which is played in a controlled environment strictly regulated by rules of the game. It can be approximated by a closed system where internal and external conflate, there is no point to distinguish between these two.

Market competition Let us imagine a few construction companies operating in a certain region. Since they operate in the same market, they compete against each other for contracts, participate in public biddings etc. Because of that, they want to keep some crucial information secret from each other. This does not mean that there is no collaboration between them: they agree on business practices, negotiate moves that would disrupt the market or form consortia to win larger contracts. In terms of our framework, external constraints and activity of the system are related to the companies construction activity and their interactions with third parties (clients, other businesses, public sector). Internal constraints regulate the game they are playing among themselves, which contains elements of both competition and cooperation, defining, for instance, which issues may be openly negotiated and which are kept secret.

4 Biological vs digital agents

In principle, the difference of natural and artificial agents should be expressible in the systemic language of degrees of freedom and constraints. Artificial and natural systems will differ in ranges of their variability and the composition of the internal and external constraints as well as the capacity of their organization in the face of interaction with the environment. Leaving, for a moment, aside the issue if it is possible to systematically distinguish the two classes of systems in terms of such patterning of degrees of freedom, we will analyze the cases of their interplay in various composite systems. Whenever two agents have different sets of external constraints there is room for synergy between them. When working together they are adding their degrees of freedom extending their joint action repertoire, and are able to act as a collective system in ways inaccessible to individuals. On the other hand, agents with overlapping sets of constraints can perform the same external activities and can act as a replacement for each other. Introducing redundancy into the system may increase its robustness to failure of individual agents. As for the internal constraints, they have to be shared (or at least compatible) between agents to facilitate coordination of their actions, allowing for "pooling" their resources in the form of respective ranges of degrees of freedom. Agents with non-overlapping set of internal constraints are usually unable to collaborate effectively. Still, in some situations limiting the communication between agents may be purposeful and beneficial. For example, we may want the medical diagnosis to be performed independently by two medical practitioners to hear possibly differing points of view. In general, separating the operation of parts of the system limits the possibility that an anomaly in one part will affect the functioning of other parts.

There are two interesting edge cases of the interplay of constraints that will help us grasp the full range of possible configuration of heterogeneous cognitive assemblages: systems in which agents have non-overlapping sets of external constraints and fully overlapping sets of internal constraints, and systems in which

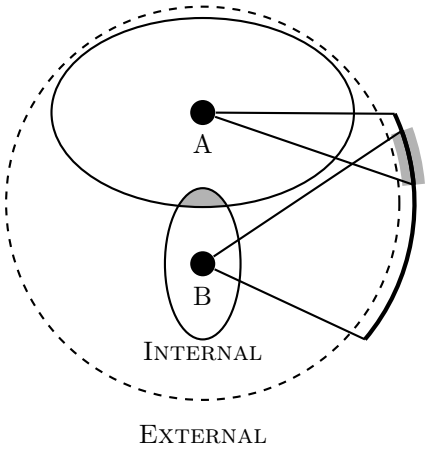


Figure 2: Specialized agents with different scopes of activity. Ellipses mark scopes of their internal activity, arcs mark scopes of their external activity. Shaded area corresponds to the amount of constraints shared between the agents. Agent A is active mostly externally, agent B is active mostly internally.

external constraints overlap and internal constraints do not. An example of a system of the first kind is a highly centralized system where agents are coordinated through a central authority and they have clearly defined roles in order to maximize their synergy and the whole system efficiency. A system of the second kind may be decentralized system in which agents act independently and may perform the same roles—such system has a lot of inherent redundancy, but may be optimized for robustness. This fits well the debates encountered within engineering and management sciences concerning merits of centralization and decentralization (Andrews, Boyne, Law, & Walker, 2007; Hugoson, 2009). Both types of systems are idealistic cases because a full separation between internal and external constraints is impossible: they always intertwine and evolve together. All realistic distributed systems are positioned somewhere along the spectrum, taking a trade-off between efficiency and robustness.

Here we can return to the issue of agents incompatibility from our broad systemic perspective. Let us begin with examples of intuitively understood "incompatibility". Consider a scenario in which two computer programs communicate using different versions of the established protocol. They communicate through discrete symbols transmitted over some channel. The code used to encode the symbols and the protocol itself impose strong bindings on the communication process: not all messages will be interpreted as meaningful at a given time (internal constraints limit the number of degrees of freedom). At some point, program A will receive a message from program B which it does not know how to interpret. At best, it can ignore the message, signal an error, and try to continue operation normally. This strategy is not likely to work in the long run as the contexts in which programs operate will get out of sync and errors will appear more often. It is to be expected that sooner or later an unrecoverable error will be encountered and both programs will crash.

Somewhat analogous example will be two people using different dialects of the same language. It is likely that not all words and expressions used by the speaker will be familiar to her interlocutor. Now, the recovery strategy will be different than in the case of computer programs. Even if some words sound unfamiliar, the receiver may guess their meaning based on similar words from her own dialect. She may also ask the speaker to rephrase his utterance using different words (binding degrees of freedom of the speaker to elicit different communicative action). Finally, she may deduce the meaning based on the situational context and the speaker's behavior.

From those two examples it is visible that the kind of incompatibility that leads to an unrecoverable error in computer programs is only a minor obstacle in human communication. This difference is due to the different number of internal degrees of freedom of the agents, different characteristics of internal constraints (very specific and limiting versus broad, flexible), and different amount of overlap in the external constraints (separated contexts of computer programs versus situational context shared by humans). While computer programs are designed in such a way as to minimize the need for communication under normal circumstances (until an exception occurs), humans willingly involve themselves in casual conversations and do so with pleasure. In fact, the very notion of incompatibility may be alien to biological systems: all organisms are connected through the shared environment (shared external constraints, i.e., being in shared, or similar "projects" as living systems (Merleau-Ponty,), and have potential of adaptation, either through individual or social learning, or through evolution (many degrees of freedom). Should there be ecological need for two organisms to mutually bind their degrees of freedom, it can be achieved through communication and/or adapting the environment to construct a commonly constraining niche. In either case a set of internal constraints providing effective communication channel or a set of external constraints regulating actions directly will develop.

Taking the presented perspective into account, we argue that incompatibility caused by cognitive heterogeneity is a problem specific to the discipline of engineering, and it stems from certain design choices. Increasing the number of degrees of freedom of the designed systems and allowing them to access broader situational context should mitigate the problem. An example of artificial system design which can cope with some level of incompatibility is a neural network. Instead of accepting discrete symbols, it operates on continuous input which is interpreted dynamically. If the input changes slightly, the network will still attempt to interpret it based on the similarity to known patterns. Adding feedback-loop (relaxing a set of constraints) would allow network to learn and adapt to even bigger changes in the input signal. Resigning from supervised learning in favor of more open-ended schema (such as reinforcement learning) might open the scope of external situations impinging systems' learning.

5 The future of artificial intelligence

Studying relations between humans and artificial intelligence was one of the motivations behind developing our modeling framework. Now we will demonstrate how some of the existing accounts of this issue may be posed in terms of the vocabulary introduced in this paper.

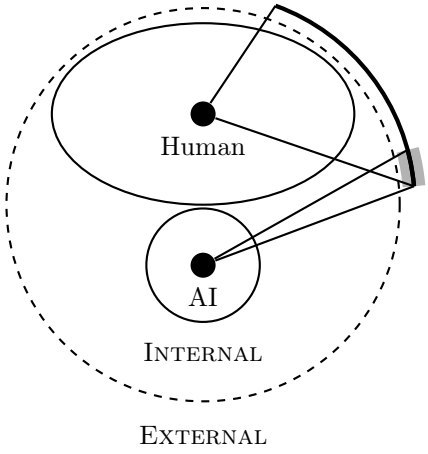
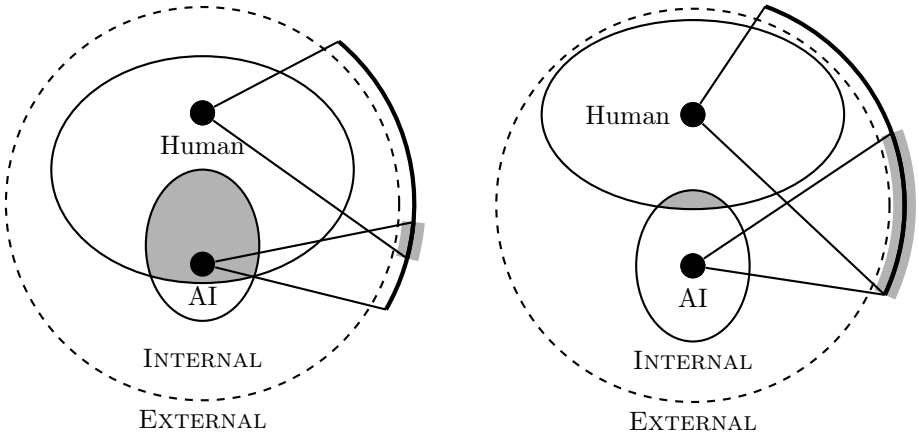


Figure 3: Human intelligence usually has more degrees of freedom and is less constrained than AI.

Luciano Floridi draws our attention to the need of proper 'enveloping' artificial agents (Floridi, 2014). By enveloping he means structuring the environment around them in such a way as to take advantage of their capabilities. For example, an industrial robot in a factory may perform the job of assembling mechanical components, but it needs to be properly orientated in space, there must be no obstacles blocking its movement etc. An intelligent dishwasher, albeit being a great convenience, still needs a human to load and unload the dishes—it can operate only inside this integrated, controlled environment. The same goes for digital programs expecting input data in a specific format and refusing to process entries with even slight formatting errors. Within our framework these observations may be expressed as follows: artificial agents generally have fewer degrees of freedom than humans. They are subjects to limiting constraints, and they can adapt to only so many changes in the environment. In order to use their capabilities within broader environment in which humans live it is necessary to artificially control the environment dynamics.

The consequences for humans coexisting with artificial agents are twofold. First, it becomes the job of a human to mediate between the enveloped AI world and everyday reality. The world needs to be quantified, digitized, and put in neatly organized databases in order to be accessible to computer systems. As these tasks are necessary performed by humans, we can argue that the proliferation of AI leads to creation of more jobs of this kind. Sometimes humans are explicitly incorporated as components of larger computing systems, as in the case of human-based computation or human-in-the-loop models (Quinn & Bederson, 2011). Second, human-made environment—digital (the Internet), social or built (for example, cities), is transformed to better accommodate for the needs of more restricted artificial agents. As Floridi argues, recent successes of AI, such as machine translations or recommendation engines, are not due to artificial agents getting smarter but due to the environment becoming better structured in the form of centralized data collections. This direction of development might be worrying because it reveals a tendency to impose strong constraints on a



a) Complementary cognitive artifact b) Competitive cognitive artifact

Figure 4: AI agents as complementary (small overlap in the set of external constraints, large overlap in the set of internal constraints) and competitive (small overlap in the set of internal constraints, large overlap in the set of external constraints) cognitive artifacts.

collective system to reduce it to its least common denominator—here the more limited repertoire of an artificial agent. Humans may be put into a constrained environment which is unnatural for them. There are serious concerns that the digitized world may be incompatible with some difficult to quantify, thoroughly human ways of relating to the world (De Jaegher, 2019).

In another take on the same issues presented by Krakauer (2016) artificial intelligence is portrayed more as a set of tools enhancing human cognition. It is rooted in the notion of 'cognitive artifact' Norman (1991). Cognitive artifacts are material objects designed to help humans with cognitive tasks. Examples would include a printed calendar, abacus, calculator etc. Krakauer introduces a distinction between complementary cognitive artifacts and competitive cognitive artifacts. The first group consists of tools which complement human cognition but do not substitute it. The way of they operate is transparent to their users, and when they are lost, the user may still perform her tasks without relying on the currently inaccessible tool. An example given is the abacus, which helps to perform arithmetic operations, and at the same time teaches useful mnemotechnics for fast calculation which may be used without an external tool. In contrast, competitive cognitive artifacts perform some functions of human cognition on their own, and their operation is often hidden from the users. When such device is lost, its user becomes even more helpless than before. Here an example may be electronic calculator which provides ready-made answers but hides the calculation process. To describe this situation in terms of our framework we need to consider external and internal activity of both cognitive artifact and its user. In the case of complementary cognitive artifacts the function of a system is distributed between both agents, their external sets of constraints do not overlap. At the same time internal activity of both agents should be rich and internal constraints should be shared, which results in the perceived transparency of operation. With competitive cognitive artifacts the situation is different: external

constraints of the artifact overlap with external constraints of its user (artifact substitutes human cognition), while their internal constraints do not. This results in the impression that the artifact is producing meaningful results but in a way which cannot be traced.

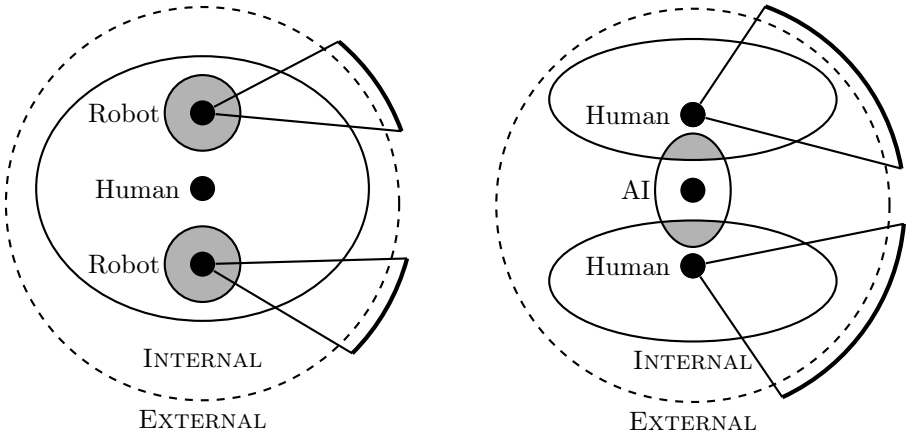
Krakauer suggests that nowadays we design a lot of cognitive artifacts of competitive nature, such as navigation systems using GPS, smart recommendation engines, etc. This may pose certain problems, as we stop relying on our own cognitive capabilities, and delegate more and more tasks to the artifacts. Such over-reliance on external tools may diminish our own abilities but also delegates the control over the situation. It is especially dangerous since people who design cognitive artifacts are usually different people than those who use them. At some point it is justifiable to ask whether we are making independent decisions as artifact users or maybe we are already controlled by artifact designers through the artifacts we use.

Analysis of the concerns of Floridi and Krakauer in terms of our framework reveals that the potential problems arise not from the nature of artificial intelligence itself but from the way the human-AI interaction is structured. Should AI have harmful impact on humanity it is mostly a consequence of bad interaction design. Some general guidelines on how to organize the relation between man and computer in a productive way were formulated at the very beginning of the digital era by Licklider (1960). He envisioned man-computer symbiosis in which humans and computers are tightly coupled in a single cognitive system possessing the best qualities of both types of agents. We describe some general possibilities for such coupling considering differences between humans and AI in terms of degrees of freedom and constraints.

Since artificial agents generally have fewer degrees of freedom than humans their role is often connected with reducing excess dynamics and decreasing complexity of the system. They can act as filters aggregating large volumes of multidimensional data, and presenting it in a form acceptable by human, for example a graphical plot. At the same they may also constrain degrees of freedom of their human partners by examining consistency and formal consequences of the hypotheses formulated by humans. This is very close to the form of cooperation originally proposed by Licklider.

An artificial agent may also play a different role. While computer programs are generally regarded as rigid and not particularly creative on their own, they may interact with a computer user in such a way to increase system's complexity, not decrease it. This is usually realized through injecting a particular kind of randomness to the system to inspire the user, and make her consider new possibilities which would normally be left unexplored. A particularly vivid example would be a recommender system suggesting books in a bookstore which from time to time makes a suggestion at random in order to broaden reader's tastes. This may be interpreted as loosening constraints and introducing more degrees of freedom.

As stated before, differences between humans and artificial agents cannot be simply reduced to different numbers of degrees of freedom. Equally important are qualitative differences between external activities of both types of agents. It is trivial to state that computers perform arithmetic operations much faster and more precisely than humans. After all, this is their original purpose as calculation machines. Computers, as opposed to humans, may operate 24h a day, do not require rest, are not affected by tiredness or moods. This makes them more



a) Human robot operator b) AI mediating communication

Figure 5: Two scenarios of mediating interaction: a) human controlling and coordinating actions of artificial agents, thus effectively acting as an intermediary between them, b) artificial agent filtering and moderating communication between humans, helping them focus on relevant dimensions.

suitable for tasks such as monitoring and controlling real-time processes, for example in a factory. With progress in robotics external actions available to artificial agents (degrees of freedom with respect to external constraints) broaden even more. For instance, we may think of specialized mobile robots which explore areas too dangerous to be accessed by humans. They can disarm mines, perform rescue operation inside a burning building etc. These are examples of systems in which humans display mostly internal activity, governing operations of the system, and artificial agents perform the actual external work.

Reversed situation, when artificial agents operate mostly within the system (internal activity), is also possible. Analyzing human behavior in natural unconstrained situation becomes very hard because of the excessive number of degrees of freedom and the large portion of context remaining uncontrolled. To limit variety of their forms of interaction it may be often desirable to filter their communication through artificial systems. In many cases analyzing an interaction of two people over a telephone or Internet chat is much easier than analyzing their face-to-face communication. If the communication device itself became an autonomous agent, it could facilitate communication in many different ways. For example, it could detect when the debate becomes too heated and order a break. In this situation artificial agents constrains human communication (introduces internal constraints) in a beneficial way regulating their behavior.

How to build a friendly and ethical AI?

The framework we are developing allows a fresh perspective on recurring issues, regarding the development of AI in such a way that it would be beneficial for the humanity in the long term. The problem of developing intelligent artificial agents which will behave in desirable way has been referred to as Friendly AI (Yudkowsky, 2008). It is a much broader endeavor than building an intelli-

gent device which is useful in one particular context. We must take into account that the artificial agent will learn and change its behavior over time, and the conditions in which its operates may change as well. Therefore, Friendly AI requires proper design of boundary conditions in which intelligent algorithm operates. We compel an aspiring AI-designer to frame this problem by asking questions concerning the system in which the agent will operate, characteristics of its external activity and the set of internal constraints shared with humans.

The first question is as follows: what is the system to which the artificial agent belongs? When we envision autonomous robots possessing some advanced form of artificial intelligence, we often ponder how they will integrate with society as a whole—a very general and abstract kind of system. It is far from how we think about humans as social beings: we assume that each individual will belong to a dozen of small-scale systems, such as family, circle of friends, team at the company she works for, citizens of a town, social group, etc. When someone functions well within her own family, there is a high chance she will also do well in other contexts. Translating these observations to artificial agents, it is practical to focus on the analysis of small systems which will constitute agent's social environment and incentives provided by this environment. It is not necessary to provide the agent with top-down social norms if it may acquire them in bottom-up fashion.

Next, we should ask, how do external constraints of the designed agent relate to external constraints of an average human individual? Since external constraints are connected with agent's actions in the world they are central to machine ethics. Moor (2006) made a distinction between implicit and explicit ethical agents. An implicit ethical agent is constrained in such a way that it avoids unethical outcomes of its actions, while explicit ethical agent is able to perform analysis of ethical categories and make ethical decisions of its own. It is an open question whether artificial agents may ever be explicitly ethical, and if so which ethical system they should adopt. Adult humans are uncontroversially considered explicitly ethical. However, in many everyday situations we do not rely on their explicit ethical analyses, but constrain their behavior to make them implicitly ethical. A simplistic example may be private property encircled by a fence which prevents unauthorized people from trespassing and invading owner's privacy. Artificial agents may by guided by the same constraints that already exist in society to constrain human behavior, provided that they share human limitations of action. An autonomous robot moving on the ground may be stopped by a fence in the same manner as any human passerby. In contrast, fence will not be an obstacle to a flying drone. Such discrepancies between AI agent and human regarding action possibilities have to be carefully analyzed as they may lead to undesirable and unexpected outcomes.

Finally, we should consider how much of agent's internal constraints are accessible to humans. AI agents displaying mostly internal activity within a heterogeneous human–AI system are less risky ethically as they do not exercise direct influence on the external world, and are better controlled by humans. For more autonomous agents acting directly in the world, sharing their internal constraints with humans is also desirable, as it provides transparency and makes decisions made by artificial agents easier to interpret. Explainable Artificial Intelligence (XAI) is an AI research sub-field advocating designing AIs in the way that makes various aspects of their creation and functioning transparent, understandable, comprehensible and interpretable by human beings (Arrieta et al.,

2019). This "opening of AI block box" is crucial to securing interests of various stakeholders in the process of introducing advanced ML algorithms to the social and economical life and an important aspect of "ethical AI" movement (Bostrom & Yudkowsky, 2014). In terms of our model, this striving to understand AIs' practical functioning translates into increasing the range of constraints shared between artificial and human components of the systems. From the human point of view it would mean the increase of the number of degrees of freedom of human component (gained from understanding how the system/AI component works) within the bounds that coupled human–AI system defines.

The mission of neurotechnology company, NeuraLink, can be thought of as an extreme version of XAI, in the sense of achieving seamless extension of the range of human degrees of freedom within the human–AI system. Recently the company announced considerable developments in hardware and neuro-surgical procedures needed to create an advanced mind-machine interface (Musk & Neuralink, 2019). Its most immediate application would be to aid patients with a number of medical conditions. However, according to NeuroLink's founder, the ultimate goal is to achieve a high-bandwidth direct interface between human minds and AI agents, which he thinks is the only way for humans to remain relevant in the age of rapid development of AI technologies (Ferris, 2017). From our perspective, NeuraLink's postualted mission is to ultimately maximise the range of degrees of freedom of human components of heterogeneous cognitive systems or, in other words, radically extend the domain of constraints shared by human and artificial agents. It remains to be seen if the final "interface" will perform the way company hopes and advertises, but we expect that our framework will be able to interpret the functioning of the final technology as appropriately as it is able to explain the initial assumptions of its development.

6 Conclusions

The world is experiencing ever increasing growth in variety of "smart" matter—we are constantly creating new classes of cognitive agents which enrich thinking ecosystem populated not so long ago almost exclusively by humans and animals. We need conceptual frameworks which would allow us to efficiently describe, model and explain the cognitive heterogeneity modern world has generated.

Because ultimately we would like to be able to model systems characterized by cognitive heterogeneity, our task was to provide a common denominator to agents' variety, but in a way that is practical—simple, accessible, preferably not creating or involving any kind of "hard problems". This is why we decided to locate this common ground in the relationship of actors and their environment, eliminating the need for direct referring to their internal processes as a source of their heterogeneity. Our approach then was pragmatic in both colloquial and, consequently, technical sense—we strived for a framework that could be easily operationalized and used in modeling, and the most straightforward way to achieve this goal is to assume that what accounts for internal cognitive architecture of an agent are the practical consequences of this architecture for agent's engagements with its environment and other agents.

This led to very rudimentary operationalizations. We were able to define dynamics of heterogeneous systems in terms of an interplay of agents' degrees of freedom and their bindings by (environmental) constraints and actions of

other agents. This very minimalistic conceptualisation resulted in a flexible and general framework which allows describing a wide range of cognitive systems and variety of processes taking place in heterogeneous systems. It remains an open framework, yet it focuses attention on often underappreciated aspects of human and human-AI functioning, such as engagement in multiple systems and projects and relation among the scope of degrees of freedom. We hope that it will help to frame the discussion on collective cognitive systems and artificial intelligence in a constructive way.

Funding

This work has been funded by Polish National Science Centre (grant number NCN Opus 16, 2018/29/B/HS1/00884 to JZ and JRL).

References

Andrews, R., Boyne, G. A., Law, J., & Walker, R. M. (2007). Centralization, Organizational Strategy, and Public Service Performance. *Journal of Public Administration Research and Theory*, *19*(1), 57–80. doi:10.1093/jopart/mum039. eprint: https://academic.oup.com/jpart/article-pdf/19/1/57/2698408/mum039.pdf

Anticipating artificial intelligence. (2016, April 28). *Nature News*, *532*(7600), 413. doi:10.1038/532413a

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2019, December 26). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *arXiv:1910.10045 [cs]*. arXiv: 1910.10045. Retrieved January 19, 2020, from http://arxiv.org/abs/1910.10045

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, *1*, 316–334.

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2018, February 28). AI Now 2017 Report. Retrieved December 19, 2018, from https://www.microsoft.com/en-us/research/publication/ai-now-2017-report/

De Jaegher, H. (2019, August 19). Loving and knowing: Reflections for an engaged epistemology. *Phenomenology and the Cognitive Sciences*. doi:10.1007/s11097-019-09634-5

Dennett, D. C. (1989). *The Intentional Stance*. MIT Press.

Di Paolo, E. A., Buhrmann, T., & Barandiaran, X. E. (2017). *Sensorimotor life: An enactive proposal*. OCLC: 982092900.

Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason* (1st ed). New York: Harper & Row.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, Mass: MIT Press.

Dworak, M., Schierl, T., Bruns, T., & Strüder, H. K. (2007, November 1). Impact of Singular Excessive Computer Game and Television Exposure on Sleep Patterns and Memory Performance of School-aged Children. *Pediatrics*, *120*(5), 978–985. doi:10.1542/peds.2007-0476. pmid: 17974734

Ferris, R. (2017, January 31). Elon musk thinks we will have to use AI this way to avoid a catastrophic future [CNBC]. Retrieved January 22, 2020, from https://www.cnbc.com/2017/01/31/elon-musk-thinks-we-will-have-to-use-ai-this-way-to-avoid-a-catastrophic-future.html

Floridi, L. (2014, June 26). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford, New York: Oxford University Press.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *32*, 147–157. Retrieved March 24, 2015, from http://www.sciencedirect.com/science/article/pii/S0732118X13000342

Hugoson, M.-Å. (2009). Centralized versus decentralized information systems. In J. Impagliazzo, T. Järvi, & P. Paju (Eds.), *History of nordic computing 2* (pp. 106–115). Berlin, Heidelberg: Springer Berlin Heidelberg.

Krakauer, D. (2016, September 6). Will A.I. Harm Us? Better to Ask How We'll Reckon With Our Hybrid Nature. Retrieved November 2, 2018, from http://nautil.us/blog/will-ai-harm-us-better-to-ask-how-well-reckon-with-our-hybrid-nature

Kuss, D. J., & Griffiths, M. D. (2017, March 17). Social Networking Sites and Addiction: Ten Lessons Learned. *International Journal of Environmental Research and Public Health*, *14*(3). doi:10.3390/ijerph14030311. pmid: 28304359

Licklider, J. C. R. (1960, March). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*(1), 4–11. doi:10.1109/THFE2.1960.4503259

Maturana, H. R., & Varela, F. J. [F. J.]. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Boston Studies in the Philosophy and History of Science. Springer Netherlands. Retrieved January 30, 2019, from //www.springer.com/us/book/9789027710154

Merleau-Ponty, M. (2002). *Phenomenology of Perception*. Psychology Press.

Miller, J. H., & Page, S. (2007, March 25). *Complex adaptive systems: An introduction to computational models of social life*. Princeton, N.J: Princeton University Press.

Moor, J. (2006, July). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, *21*(4), 18–21. doi:10.1109/MIS.2006.80

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 555–572). Synthese Library. doi:10.1007/978-3-319-26485-1_33

Musk, E., & Neuralink. (2019, August 2). An integrated brain-machine interface platform with thousands of channels. *bioRxiv*, 703801. doi:10.1101/703801

Norman, D. A. (1991). Designing Interaction. In J. M. Carroll (Ed.), (pp. 17–38). New York, NY, USA: Cambridge University Press. Retrieved November 2, 2018, from http://dl.acm.org/citation.cfm?id=120352.120354

Pattee, H. H. (1972). The nature of hierarchical controls in living matter. In *Foundations of Mathematical Biology* (pp. 1–22). Academic Press. Retrieved April 11, 2016, from http://www.sciencedirect.com/science/article/pii/B9780125972017500085

Quinn, A. J., & Bederson, B. B. (2011). Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Confer-*

ence on Human Factors in Computing Systems (pp. 1403–1412). CHI '11. doi:10.1145/1978942.1979148

Riley, M. A., Richardson, M. J., Shockley, K., & Ramenzoni, V. C. (2011). Interpersonal synergies. *Frontiers in Psychology*, *2*. doi:10.3389/fpsyg.2011.00038

Riley, M. A., Shockley, K., & Van Orden, G. (2012, January). Learning from the body about the mind. *Topics in Cognitive Science*, *4*(1), 21–34. doi:10.1111/j.1756-8765.2011.01163.x

Searle, J. R. (1980, September). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. doi:10.1017/S0140525X00005756

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Varela, F. J. [Francisco J.], Rosch, E., & Thompson, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, Mass: MIT Press.

von Bertalanffy, L. (1968). *General system theory: Foundations, development, applications*. G. Braziller.

Yudkowsky, E. (2008). Cognitive biases potentially affecting judgment of global risks. *Global catastrophic risks*, *1*(86), 13.

Zelený, M., & Hufford, K. D. (1992, July 1). The Application of Autopoiesis in Systems Analysis: Are Autopoietic Systems Also Social Systems? *International Journal of General Systems*, *21*(2), 145–160. doi:10.1080/03081079208945066