The structure of intentional action: An information-theoretic account

Stipe Pandžić[1], Jan Broersen[2], and Henk Aarts[3]

[1]LUCI Lab, University of Milan, Italy
[2]Department of Philosophy and Religious Studies, Utrecht University, The Netherlands
[3]Department of Psychology, Utrecht University, The Netherlands
stipepandzic@gmail.com

Abstract

We propose a five-dimensional model for intentional action based on information theory. Building on neuroscience models of intention, we select five key features of intentional action: conflict level, cognitive control, memory, endogeny, and pattern complexity. We argue that these five features constitute hallmarks of the structure of intentions. Using information-theoretic tools, the five features are expressed as generalized measures relating agents to their environment and (past) actions. The resulting ensemble of measures serves as a blueprint for an operational definition of intentional action applicable to both artificial and biological agents.

Submission type: **Full Paper**

Introduction

Most philosophical theories of intentional action require the attribution of mental properties and the presupposition of a first-person perspective for agents. Philosophers characterize intentional action, for instance, in terms of internalized reasons directing action, goals that motivate action, or simply as having the intention to perform an action. Anscombe (1963, § 6) suggests that "intentional actions are those to which a certain sense of the question 'why ?' has application". In such views, explaining what intentional action is requires the folk-psychological language of internal mental states, famously criticized by Dennett in his intentional stance view Dennett (1989).

In this paper, we want to characterize intentional action in terms of the observable characteristics of agency, thereby sidestepping the vexed issues of investigating intentional action as an element of 'the realm of the mental'. We use tools from information theory to demarcate the *intentional* in the behavior of agents. Such an 'externalist' view on intentional action helps us to make the concept of intention operational, which has proven to be challenging on internalist accounts of intention.

At the core of this proposal is the idea that *recognizing* intention is an important aspect of social cognition and that the advent of autonomous agents makes intention recognition a more challenging task and a pressing societal issue. The question we want to address is 'how to recognize whether an action that an agent undertakes is intentional or not?', using only *observable* data about an agent and its environment.

Humans need to engage in intention recognition by inferring hidden goals and motives to predict and adapt to the actions of their peers. Researchers have attributed this intention recognition mechanism to the activation of the mirror neuron system (Iacoboni et al., 2005). However, this theory depends on the mechanism of intra-species simulation and on the idea of similarity between agents. It is unlikely that such intention recognition mechanisms can be easily applied to computational models of intention detection, in particular for intentions of agents that are potentially very different from humans. The issue of recognizing intentions of non-human agents has recently been raised by philosophers in the context of the large-scale use of generative artificial agents and AI-directed drones (Martela, 2025, p. 4390).

Our operational definition of intentional action aims to characterize intention as a matter of degree rather than 'intention simpliciter'. This is in line with human practices of intention attribution, where certain domains explicitly require establishing a level of intent. A good example is criminal law where courts need to assess the extent to which an action counts as intentional, which affects eventual attributions of criminal responsibility.

We expect that our model can play a role in tackling the notorious alignment problem Ji et al. (2025); Hägele et al. (2026) in AI. In alignment problems in AI, it is often not made explicit what should align with what, but implicitly it seems most often assumed that what should be aligned between humans and AIs are their intentions. We think the 'structural view' of intentional action that we advocate for below may be suited to analyze Human-AI alignment.

Before proceeding, we point to the distinction between intentional action and intention as a state, which gained prominence in the mainstream analytic philosophy of action (see, e.g., Searle (1980) and Bratman (1984)). There is a difference in detecting intentionality as 'directedness' and detecting intentions as 'directedness toward a goal'. Our definitions

remain by and large agnostic with respect to the *content* of goals and plans, but they still aim to capture the second idea of being directed toward a goal. So, we partially subscribe to the 'Simple View' (Bratman, 1984, p. 377) on the above distinction, that is, the view that an action α is intentional if α is appropriately related to a state of 'intention to α'. In the next sections, we generally assume that an intentional action α and an intention to α are appropriately related, but without assuming any properties of intentions as 'mental states'.[1]

The paper is structured as follows. We first select five key features of intentional action based on neuroscientific models. After briefly justifying our methodological choices, we present the five-dimensional quantitative account of intentions and interpret the resulting definition of intentional action as a set of directed dimensions. Finally, before concluding, we discuss related work and point out several limitations of the current framework.

Selecting key features of intentional action

We propose a five-dimensional account based on recent neuroscientific models of intentional action. The first dimension concerns the resolution of response conflicts as a mechanism of intentional action. This aspect emphasizes that any intentional action is only one of the competing responses to actual stimuli. Competition between the available options is stronger when an action is not simply a response to an associated triggered stimulus. At the extreme, competition seems stronger when an action results from a "free selection situation" because "all potential response alternatives have a rather similar activation level" (Brass and Haggard, 2008, p. 320).

The presence of competing alternatives necessitates a mechanism that selects behavior which is conducive to an agent's goal. This mechanism is known as 'cognitive control' that could be defined as those processes that "refer to our ability to coordinate thoughts and actions in accordance with internal goals" Brass et al. (2005). Cognitive control can be understood as an ability to suppress distractor cues in the environment to prioritize goal-oriented behavior. The second dimension is thus one of the ways to capture the commitment aspect of intentions, specifically as commitment against the environment in which the action takes place. This dimension of intentional action is unique because it operates "online by intervening between stimulus-driven processes and the production of actions" Memelink and Hommel (2013). We associate a local (event-specific) measure with this dimension, reflecting its uniquely online control mechanism.

The third dimension takes into account the fact that intentional action needs to be mediated by memory, rather than resulting from direct stimulation.[2] That is, we need a measure that approximates the prospective memory or the memory for intentions (Einstein and McDaniel, 1990). This dimension is important because it will also serve as a surrogate measure for detecting the goal-directedness or the teleological component of intentional action, while relying solely on observable patterns in behavior.

The fourth dimension concerns the expectation that intentional action does not trivially connect to the environment. This stands in contrast to stimulus-driven responses and reactive behavior, for which trivial connection is a defining characteristic. Although agents form intentions at least partially in reaction to their environment, the connection to the environment is looser than in the case of stimulus-driven action (Brass and Haggard, 2008, p. 321). In general, we do not expect that a single external event is fully predictive of what action an agent undertakes next.

Although we do not expect intentional action to follow immediately after an external stimulus, we do assume that the environment meaningfully influences the agent's choice of which action to undertake next. This leads us to include another related dimension motivated by the neuroscience perspective on the complexity of generating intentions. The fifth dimension is based on the finding that intentional action is related to "complex contextual patterns rather than identifiable single stimuli" (Brass and Haggard, 2008, p. 321). We take this to not only imply that in deciding on undertaking an action, an agent uses a relatively high number of observations but also that the agent's intention reflects some state of affairs taken as a 'whole'.

We note that none of the above mentioned aspects counts as a stand-alone indicator of intentional agency. For example, the fact that an action is just one among many conflicting responses to an external stimuli does not mean that the action is intended in any way. To see this, we can simply assume that an agent has a great number of action choices and throws a die to pick one action. This type of uncertainty is usually not a token of intentional choice, unless there is a higher-level intention to be surprised.[3] There are also multiple conceptual connections that relate the features of intentional action discussed above. Thus an action may be mediated via memory traces, e.g. when it is a part of an elaborate plan to attain a goal, and memory plays an important role in processing com-

[1] Another well-known philosophical distinction on intentions is that between 'original/intrinsic' and 'derived' intentionality Searle (1983); Dennett (1989). Original intentionality is human intentionality, and derived intentionality is all the intentionality that derives from it (for instance, in our texts and in the machines we design, including artificial agents). For our externalist approach to the characterization of intentionality, the distinction is not that important.

[2] This does not mean that memory is not important for stimulus-based action. The difference is that intention-based actions and stimulus-based actions rely on different memory traces in such a way that "the activity of the system guiding stimulus-based actions is accompanied by stimulus–response (sensorimotor) learning, whereas the activity of the system controlling intention-based actions results in action–effect (ideomotor) learning" (Herwig et al., 2007, p. 1549).

[3] In the latter case, the relevant intentional structure lies at a different level and shifts toward the goal of being surprised.

plex pattern in the agent's environment. Another connection is that between cognitive control as a system of shielding an internally adopted goal against distractors and resolving response conflicts as a way of selecting an intended action.

Methodology

We use tools from information theory to detect the presence of intentions in the behavior of agents. Recent developments in complex systems science position information theory as a promising unifying framework in the philosophy of mind. Successful applications include the concepts of consciousness Tononi et al. (2016), autonomy Bertschinger et al. (2008) action-perception cycle Buckley et al. (2017), and emergent phenomena in general (Varley and Hoel, 2022; Prokopenko et al., 2009). We propose to operationalize *intentions* as an ensemble of information-theoretic measures that encodes the structure of the interaction between action and environment.

We will not rely on sub-symbolic methods to analyze data in search of intentional action patterns, because we are not simply interested in the 'predictive' patterns about a specific type of agency. The starkest downside of sub-symbolic methods is that, beyond finding local regularities, they do not uncover broader principles. Our aim is to follow information on a more abstract level to generalize the outcome measures to any type of agent and agency. Although we learn from neuroscience models of intentions, we seek to give a *universal* rather than an *anthropomorphic* characterization of intention.

Instead of identifying intentions by recognizing particular action patterns, we propose detecting whether an action can be explained as intentional based exclusively on the *structure of information* or 'information traces' that any such pattern would produce. Information traces include different information-theoretic measures of, e.g., the history that led to a specific action of choice, available alternatives to the chosen action, the level of complexity of the contextual patterns that is predictive of the current action choice, and so on.

Methodologically, our focus on information traces of actions sets a strong behaviorist tone. This will become clear in the following section, where we restrict the definitions to the interplay between actions and environment. We deliberately choose to investigate the potential of such a reductionist approach to accommodate different types of agents, be it biological or artificial.

An information-theoretic account of the five-factor intention model

In this section, we give the five information-theoretic measures for the concepts introduced in the previous section. Note that we are starting from a simple setting of contrasting actions and the environment in which an action takes place. There are n time steps and at each time step t, the agent receives an input multiset of events $E_t = \{E_{t,1}, E_{t,2}, \ldots, E_{t,n}\}$ and takes an action A_{t+1}. Each event variable $E_{t,i}$, for $1 \leq i \leq n$, takes values from $\{e_1, \ldots, e_m\}$, and each action A_t takes values from $\{\alpha_1, \ldots, \alpha_k\}$ This gives jointly distributed random variables:

$$E_{1,1}, \ldots, E_{1,j}, \ldots, E_{t,1}, \ldots, E_{t,k} \quad \text{and} \quad A_1, \ldots, A_n.$$

In general, the following may hold $n \neq m \neq j \neq k$. The set of event multisets generated by the environment over the finite number t of time steps is defined as $\mathbb{E} = \{E_1, E_2, \ldots, E_t\}$. For the resulting discrete-time interaction process indexed by $1, 2, \ldots, t$, the sequence $(E_{1,1}, \ldots, E_{1,i}, A_1, E_{2,1}, \ldots, E_{2,j}, A_2 \ldots, E_{t,1}, \ldots, E_{t,k}, A_t)$ is assumed to be jointly distributed according to an unknown but stationary probability distribution P. Note also that the joint distribution between these random variables is not completely arbitrary. In particular, a number of temporal constraints limit the joint distribution:

- A_t cannot depend on future inputs $E_{t+1}, E_{t+2}, \ldots$;
- E_t cannot depend on future actions $A_{t+1}, A_{t+2}, \ldots$;
- A_t can only depend on past inputs $E_1, \ldots, E_{t-1}$;
- A_t can only depend on past actions $A_1, \ldots, A_{t-1}$.

No other assumptions are made at this point, with the aim of keeping the definitions simple and flexible to capture the original motivation drawing from neuroscience models.

Conflict level via conditional entropy The first dimension of the entropy-based measure for intentional action concerns the condition of having relevant alternatives to the actions that are being executed by an agent. Such alternatives matter because they witness that there is an expectation of multiple possible responses to an agent's environment. On such views (Nachev et al., 2005), intentional action is a way to resolve response conflicts in selecting an action over its competing alternatives.

The analogy between the information theory concept of uncertainty and the behavior theory concept of conflict has already been recognized in the literature, e.g., by Berlyne (1957). He proposes that uncertainty can be seen as an indication of the complexity of a conflict. Such perspective reduces conflict to the difficulty of predicting which of the conflicting responses will be the one to occur. This is in line with our information-theoretic approach to defining intentional action. An additional aspect to our measure will be the context generated by the environment $\mathbb{E}$. Thus, the measure for the level of conflict will be the measure of uncertainty relative to the context, which is captured by *conditional 'Shannon entropy'* (Shannon, 1948, p. 395) with the joint random variable $\mathbf{E}_{t,1\ldots n}$ in the condition:

$$H(A_{t+1} \mid \mathbf{E}_{t,1\ldots n}) = H(A_{t+1} \mid E_{t,1}, E_{t,2}, \ldots, E_{t,n}) = - \sum_{a_{t+1}, e_{t,1}, \ldots, e_{t,n}} p(a_{t+1}, e_{t,1}, \ldots, e_{t,n}) \log p(a_{t+1} \mid e_{t,1}, \ldots, e_{t,n})$$

where $p(\cdot)$ denotes the appropriate joint and conditional probability mass functions derived from the underlying distribution, and the logarithmic base 2 gives the output measured in bits. The analogy between the level of conflict and the entropy level is straightforward. Namely, a higher level of conflict corresponds to a higher measure of entropy.

Cognitive control via local misinformation A feature of intentional action that is closely related to that of conflict monitoring is cognitive control. Cognitive control can be defined as the ability to "orchestrate thought and action in accordance with internal goals" (Miller and Cohen, 2001, p. 167) and it has been recognized as an essential ability for intelligent behavior (Miller, 2000). One of the proposed functions of cognitive control is reducing uncertainty, which intuitively connects cognitive control with Shannon entropy. This intuition motivated Mackie et al. (2013) to examine cognitive control from an information theory perspective.[4]

In our interpretation of intentions, we focus on the discrepancy between probable and actual responses. This aspect has been emphasized by Alexander and Brown (2011, p. 1338) who suggest that cognitive control should be "seen as a result of evaluating the probable and actual outcomes of one's actions". We are interested in framing the dimension of cognitive control characteristic of agents who favor to undertake a goal-oriented action, despite having distracting cues in their environment. We will use information theory tools to quantify the level to which an action that an agent eventually undertakes has been contested by distractor cues in its environment, such that these distractor cues lessened the probability that the undertaken action would have taken place.

More specifically, we will use *misinformation* or a measure of negative pointwise information terms. Misinformation has been recognized as a significant quantity in the area of prediction errors and information processing in the brain (Ince (2017), Wibral et al. (2014)). It concerns negative measures of local mutual information for a pair of variables. In contrast to positive values of local mutual information where knowing the measurement value y increased the expectation of the measurement x, negative values mean that knowing y decreased the expectation of x, even though x occurred. Therefore, knowing y has been interpreted as misinformative in this context (Wibral et al. (2014), Ince (2017)).

We use misinformation to measure the level of possible interactions that may misdirect a system acting within its environment, having in mind a background assumption that the environment is represented through the system's observations. The (cumulative) level of misinformation is then interpreted as measuring the *resistance* to misleading cues that result from the interaction between the system and its environment. The measure of 'resistance' indicates whether a selected action is carried out against competing predicted responses that could have been triggered in a given state of the environment.

The concept of 'misinformation' measure is based on the possibility of negative values for *local mutual information*:

$$i(\alpha; e) = \log_2 \frac{p(\alpha, e)}{p(\alpha)p(e)} = \log_2 \frac{p(\alpha \mid e)}{p(\alpha)},$$

$$\text{where } p(\alpha \mid e) < p(\alpha).$$

That is, e lowers the probability $p(\alpha \mid e)$ below the initial probability $p(\alpha)$, but α still occurs. Intuitively, we are less likely to see the two particular outcome values α and e together, than it would be expected if they were independent. This formalizes the expectation that the (perceived) features of the environment are relevant to how likely is it that some behavior will occur in the sense that some variable e might misdirect our expectations of the course of actions.

Cognitive control mechanisms must deal with stimuli that have the potential to update goal representations. For example, on your way to work, you might notice a 20$ bill lying on the ground (Miller and Cohen, 2001, p. 187). Although the probability distribution of the 20$ whereabouts variable is of neutral predictive value with respect to the course of actions of an agent, this information may be locally relevant. It is this type of distracting cue that might result in straying away from one's initial goal, for example, if you actually stop to pick up the bill.

The cognitive control measure is the only measure in which we look at *local*, event-specific properties, instead of distribution-level properties. We deliberately combine local and global measures because localized signals often need to be overridden to ensure stable execution of an intention. This dimension of intentions is not reflected in global measures that primarily capture structural instead of situational properties. However, to use the local values in the ensemble, we work with the cumulative misinformation measure for an action $\alpha \in A$ and a multiset $\mathbb{E}$:

$$i_\Sigma(\alpha; \mathbb{E}) = -\sum_{E_k \in \mathbb{E}} \sum_{e_{k,n} \in E_k} i(\alpha; e_{k,n}),$$

where $p(\alpha \mid e_{k,n}) < p(\alpha)$. The measure can be interpreted as the cumulative uncertainty in the expectation that the action α took place in relation to the state of the environment. We expect goal-directed systems to carry through with their actions despite the uncertainty of the outcome that is introduced by a number of distractor cues in their environment.

Memory via information storage Unlike stimulus-based responses, intentional action is mediated via memory traces (Brass and Haggard, 2008, p. 321). Research on the role of memory in human subjects showcases the importance of prospective memory or memory for intentions in carrying out

[4] Fan (2014) gives an information theoretic account of cognitive control. Fan (2014, p. 8) argues that "the ultimate goal or the function of cognitive control is to reduce uncertainty".

planned actions. In retrieving intentions, prospective memory is needed to remember when and what has to be done.

Our framework builds on observable properties of agency and it, therefore, allows only for an indirect formalization of memory in intentional action. To approximate this component, we use the measure of *information storage* that reflects the degree of information preservation in an evolving system (Xiong et al., 2017, p. 4). For a history of a process A defined as a series of actions ($\mathbf{A}_{\overline{n}}$) at the present time n or

$$\mathbf{A}_{\overline{n}} = (A_1, \ldots, A_{n-2}, A_{n-1}),$$

we define the 'memory' $M(A)$ of A in terms of the mutual information between A_n and the joint random variable $A_{\overline{n}}$:

$$M(A) = MI(A_n; \mathbf{A}_{\overline{n}}) = \sum \left[\log_2 \frac{p(\alpha_1, \ldots, \alpha_n)}{p(\alpha_1, \ldots, \alpha_{n-1})p(\alpha_n)} \right].$$

We use the measure of memory $M(A)$ to detect whether an agent's behavior shows that information flows from past to future actions. In contrast to $M(A)$, or *memory for intentions*, memory for habits requires the coupling of actions and the environment, that is, a higher value of $MI(A_n; E_{n-1,1}, \ldots, E_{n-1,k})$.[5] Intentional commitment requires $M(A)$ to be high even when the latter coupling of actions and the environment is low. This, in turn, can be taken to reflect the dominant use of memory for future plans and intentions in selecting actions.

Note that memory is the only component in which we explicitly consider intentions as a process that evolves over time and commits agents to a time-frame for their execution. The memory component is, therefore, our way to capture intentions as commitment toward a goal Georgeff and Rao (1991) and intentions as commitment toward time van Zee et al. (2020), but without using the language of mental states typical for *Belief-Desire-Intention* (BDI) architectures.

As a result, information storage or the memory component becomes central both to the current theoretical framework for intentional action and to the design of an intention-detection algorithm that will operationalize our account. The importance of this component lies in its potential to act as a proxy for detecting goal-oriented sequences of actions, which is the paramount marker of intentions.

[5]Undoubtedly, goal-directed reasoning also depends on the environment, and this is reflected in the expectation that intentional action depends on complex patterns in the environment. The difference is that, in the context of habitual behavior, $MI(A_n; E_{n-1,1}, \ldots, E_{n-1,k})$ is high, but due to the higher value of redundant information and the relatively low value of synergistic information compared to the synergy levels involved in goal-directed action. This standard, but contested, view of habits can be summarized as the following claim (Gardner et al., 2023, p. 522): "a 'context' tends to be used to denote real-world settings that incorporate multiple cues, while a 'cue' is typically used to denote a specific stimulus representing one of many potential lower-order fragments within a higher-order 'context' ".

Endogeny via conditional mutual information However, it is still possible that such patterns do not come as a result of memory patterns, but rather as a result of the regularities that exist in an agent's environment. Thus any plausible account of intentions should accommodate keeping track of whether an action is 'exogenous', that is, controlled by an external cause or 'endogenous' or comes from an agent. In modeling human intentional action as endogenous, we require that there is a looser connection to external events, in contrast to stimulus-driven actions (Brass and Haggard, 2008, p. 321).

By adopting a behaviorist outlook, we are prevented from claiming that any action could truly be internally selected. However, our definition must recognize that intentional action conditioned by the agent's environment differs from reactive behavior driven by the environment. This intuition is grounded in neuroimaging findings with the conclusion "that externally guided actions are generally less complex than internally guided actions" (Krieghoff et al., 2009, p. 2). One way to eliminate the option that an action is simply reactive is to confirm that the action is not a response to an immediate external trigger, i.e., an event in the environment that could be seen as a 'cause'.

This requirement will be formalized as a measure of 'decoupling' of actions from the individual events in the environment. A system is seen as (trivially) decoupled from the environment when there is no information flow from the environment to the system. We do not require such strong decoupling, but rather that the information flow from each single variable in the environment is minimal. We follow Bertschinger et al. (2008) in their formalization of decoupling via *conditional mutual information*, which they see as a precondition for autonomous agency.[6] An agent's actions are closed form (single) events in the environment when

$$\begin{aligned} &\forall i \forall k E_{i,k} \in E_i \text{ such that} \\ E_i = \{E_{i,1}, \ldots, E_{i,n}\}, &1 \leq k \leq n \text{ and } E_i \in \mathbb{E}, \\ &\text{it holds that } MI(A_{i+1}; E_{i,k} \mid A_i) \approx 0, \end{aligned}$$

where $MI(A_{i+1}; E_{i,k} \mid A_i)$ is conditional mutual information. The resulting equation can be interpreted as a way to determine the added amount of information that an individual event provides about the next action, given the previous action. Our requirement is that none of the events, taken on its own, should be predictive of an action choice, if this action is intentional. This condition does not imply independence from the environment as a whole, only from any single environmental variable.

[6]Note that Bertschinger et al. (2008, p. 334) use the nonstandard notion of information flow (IF). They give an insightful information-theoretic account of autonomy. The subject areas of autonomous agency and intentional agency are interwoven and it is not surprising that their features overlap to some extent. It seems that the experience of intentional agency figures as a key component of personal autonomy (Antusch et al., 2021), but the exact relation between the two is intricate.

Complex patterns via synergistic information Instead of being merely reactive, we expect that intentional action is based on complex contextual patterns. The information closure measure above decouples intentional action from single stimuli, but intentional action should be based on interacting with the environment and using observations about the context to achieve context-related goals. Thus, our intention detection procedure needs to include a component that measures the extent to which complex features of an agent's environment can be used to predict the next actions of the agent. Although the measure of endogeny decouples single events from the choice of an action, intentional action is not independent of the environment. However, this dependence is not trivial.

Both pre-theoretical considerations and modern neuroscience (Brass and Haggard, 2008) converge on the idea that intentions build from complex contextual patterns. We take this idea to impose the requirement that intentional action needs to result from a wider whole of the (observed) events in the environment. This requirement brings about a measure that can be seen as a 'flipside' of the above endogeny measure. That is, we now search for information that is present in the events when taken together, but not in any singleton predictor.

We propose to define this dependence via a measure of *synergistic information*, which is usually understood as information carried by 'the whole' beyond the information stored in its parts (Griffith and Koch, 2014, p. 164). A typical example of synergistic information is the XOR gate. The XOR gate implements the so-called 'exclusive disjunction'. That is, the XOR gate returns a true output if, and only if, either of the two inputs to the gate is true, but not both of them. Thus, the information that the XOR gate provides is only specified by taking both inputs together, but none of the inputs by themselves carries this information.

Among the currently available synergy measures, e.g., (Quax et al. (2017), Williams and Beer (2010), Rosas et al. (2020)), we settle on the one proposed by Griffith and Koch (2014). Griffith and Koch (2014) focus on higher-order interactions among variables and, importantly, generalize the measure to arbitrarily many sources.[7] Synergistic information is defined as the mutual information contained in the whole minus the union information as follows:

$$S(\bigcup_{i=1}^{n}\bigcup_{k=1}^{s(i)}\{E_{i,k}\}; A_{n+1}) = MI(\mathbf{E}_{1\ldots n,1\ldots s(\cdot)}; A_{n+1}) - MI_U(\bigcup_{i=1}^{n}\bigcup_{k=1}^{s(i)}\{E_{i,k}\}; A_{n+1}),$$

where $\mathbf{E}_{1\ldots n,1\ldots s(\cdot)}$ is the joint random variable of all atomic predictors $E_{i,k}$ for $1 \leq i \leq n, 1 \leq k \leq s(i)$, such that $E_i = \{E_{i,1}, \ldots, E_{i,s(i)}\}$ is a set of atomic predictors of size $s(i)$ and $E_i \in \mathbb{E}$.[8]

Paired with the endogeny measure, this measure adequately captures the contrast between union or single-predictor information and synergistic information that aligns with the contrast between single stimuli and complex contextual patterns, while staying within the general partial information decomposition framework. The problem with the formulation above is how to interpret the union information $MI_U(\bigcup_{i=1}^{n}\bigcup_{k=1}^{s(i)}\{E_{i,k}\}; A_{n+1})$, since there is no established analytic measure of the 'union information'. Griffith and Koch (2014, p. 171) propose the following measure:[9]

$$MI_U(\bigcup_{i=1}^{n}\bigcup_{k=1}^{s(i)}\{E_{i,k}\}; A_{n+1}) = \min_{\substack{P^*(E_{1,1},\ldots,E_{1,s(1)},\ldots,\\ E_{n,1},\ldots,E_{n,s(n)},A_{n+1})}} MI^*(\mathbf{E}_{1\ldots n,1\ldots s(\cdot)}; A_{n+1}),$$

where $P^*(E_{i,k}, A_{n+1}) = P(E_{i,k}, A_{n+1}),$ for each i, k and

$$MI^*(\mathbf{E}_{1\ldots n,1\ldots s(\cdot)}; A_{n+1}) = D_{KL}[P^*(\mathbf{E}_{1\ldots n,1\ldots s(\cdot)}, A_{n+1}) \parallel P^*(\mathbf{E}_{1\ldots n,1\ldots s(\cdot)})P^*(A_{n+1})].$$

The Kullback-Leibler divergence tells us the distance between the joint distribution of the predictors and the target variable, under the modified probability P^*, and the product of its marginals, i.e., the distribution obtained by assuming that the predictors and the target are independent.

For intentional action, detecting synergy can be seen both as a marker of context dependence and as a measure of the informational complexity underlying intentional states. Beyond detecting and quantifying synergy as such, an interesting direction for information-theoretic measures for intentions and goals opens up by considering the importance of higher-order types of interactions beyond pairwise dependencies Combrisson et al. (2025). Although we believe that the measure S captures the idea of non-trivial dependence of intentions on the environment, it remains an open question whether identifying orders of interactions beyond pairwise might also discriminate between different goal types.

Intentional action as geometric region in information space

Taking stock of our analysis, our main claim is that the measures represent five projections of an underlying intentional structure. When viewed through the lens of the five measures, two main characteristics of intentions emerge. First, the five-factor account is expected to provide a set of directed dimensions that, taken together as an *ensemble*, allow us to

[7] The proposed measure is compatible with partial information decomposition, but it does not provide a complete decomposition.

[8] The set $\bigcup_{i=1}^{n}\bigcup_{k=1}^{s(i)}\{E_{i,k}\}$ is the set of all atomic predictors $\{E_{1,1}, \ldots, E_{1,s(1)}, \ldots, E_{n,1}, \ldots, E_{n,s(n)}\}$.

[9] In the original notation of Griffith and Koch (2014, p. 171), 'MI_U' is denoted by 'I_{VK}' after "Virgil-Koch".

assess whether an action qualifies as a genuine intention. Importantly, none of the five features alone provides a sufficient condition for intentional action. Secondly, the resulting definition specifies an action as intentional as a matter of *gradual measure*, rather than a binary property.

Accordingly, we define intentional action as occupying a region in information space where 1.) action entropy is (relatively) high; 2.) multiple misinformative outcomes have been detected in the environment; 3.) past actions are informative about the next action; 4.) no single predictor reliably predicts the next action; and 5.) the environment predictors synergistically inform the next action. Table 1 summarizes all the structural properties of intentional action and connects each of the properties to an appropriate measure.

Measure	Value		Interpretation
$H(A_{t+1} \mid \mathbf{E}_{t,1\ldots s(t)})$	$+$	$\rightarrow$	*conflict*
$i(\alpha; e'), i(\alpha; e''), \ldots, i(\alpha; e^{(\cdot)})$	$-$	$\rightarrow$	*control*
$MI(A_n; \mathbf{A}_{\overline{n}})$	$+$	$\rightarrow$	*memory*
$MI(A_{i+1}; E_{i,k} \mid A_i)$	≈ 0	$\rightarrow$	*endogeny*
$S(\bigcup_{i=1}^{n} \bigcup_{k=1}^{s(i)} \{E_{i,k}\}; A_{n+1})$	$++$	$\rightarrow$	*complexity*

Table 1: Information-theoretic dimensions of the structural definition of intentions

The resulting definition combines global and local measures. This combination reflects the twofold nature of a phenomenon attributed primarily to stable and coherent structures but nonetheless context-sensitive and enacted locally in (and often 'against') the environment. Philosophically, the proposed account is an exercise in how to operationalize intention as a commitment, that is, a mode of control over actions, without invoking goal representations. The resulting measure captures a shift from Anscombe's 'why?' question to our 'how?' question for the detection of intentions.

Related work

Our work shares its main philosophical tenets with the program of formalizing Dennett's intentional stance presented in, e.g. McGregor and Chrisley (2020) and McGregor et al. (2024). One of the key aspects of the formal account of intentional stance is integration of information-theoretic tools to model the observer or 'theorist' as adopting the 'intentional stance' whenever normative-epistemic states that correspond to belief-like information and goal-directed structures offer predictively accurate and compressed representations of behavior. Consequently, goal attribution depends on whether the variables that correspond to 'beliefs' and 'goals' provide an efficient explanatory model of observed regularities. Technically, the system's behavior can be interpreted from the intentional stance if it can be explained as 'acting optimally' in order to reach some goal (McGregor et al., 2024, p. 6). This project is philosophically promising and it could avoid some of the potential pitfalls of our framework. For example, intentional descriptions are regarded as a stance or perspective taken by a theorist, rather than an inherent property of a system (McGregor et al., 2024, p. 2).

Models for intention recognition are often based on applying Bayesian networks. For example, Han (2013) use Bayesian network inference of the intention performed by a probabilistic logic system. In Orseau et al. (2018), the authors contrast agents and devices, where agents are understood as systems that optimize a utility function, and devices follow a simple input-output mapping. This Bayesian framework is flexible to accommodate agents changing goals along their trajectory (as long as shifting goals is relatively rare), and the authors present clear experimental confirmation of their framework. The major difference from our system is that we do not rely on goal assumptions. Among the approaches that are not based on Bayesian networks, Zhang et al. (2023) aim to generalize their intention recognition techniques to multiple agents. Their system is based on the identification of 'landmark' states in the sequences of actions, where a landmark state signals intermediate achievements of agents that are obtained before completing the main task.

Bonchek-Dokow and Kaminka (2014) define a state-distance measure that determines the optimal sequence of actions between two states of the world, given all possible actions to reach one state from another. Interestingly, Bonchek-Dokow and Kaminka (2014, p. 64) find a negative correlation between the entropy and the intention when comparing 'intention graphs' with 'entropy graphs'. This means that the decrease in entropy levels can be interpreted as an increase in structure brought about by goal-oriented or intentional action. A similar expectation can be based on our measure of information storage, given that this measure can be expressed using the relation between *Kolmogorov-Sinai measure* of entropy rate, entropy and information storage (Xiong et al., 2017, p. 8).

To our knowledge, few approaches avoid externally assigning a system's goal and still address goal-oriented behavior. One such system is proposed by Kolchinsky and Wolpert (2018), who define semantic information as the syntactic information (or statistical correlation) that a system has about its environment which is causally necessary for the system to stay in low-entropy and maintain itself. The only assumption made is that a system has an intrinsic 'goal' of self-maintenance. Although viability and negative entropy cannot capture all possible goals, the framework that Kolchinsky and Wolpert (2018) develop is interesting, as it presents goal-oriented behavior as an emergent property. Clearly, the focus on semantic information differs from our focus on structural information. The work of Kolchinsky and Wolpert (2018) shares some methodological assumptions in the detection of agents with the approach of Kenton et al. (2022). In particular, both frameworks use counterfactual analysis and causality, but they diverge on what constitutes a 'goal'. Kenton et al. (2022) treat goals as discoverable from any set

of variables. Although their definition is more encompassing, it is observer-relative in the sense that the outcome of goal ascription depends crucially on which set of variables is chosen for counterfactual analysis.

One of the approaches that is methodologically closest to ours is the theory of individuality proposed by Krakauer et al. (2020). Although there is no mention of the goals of the system in Krakauer et al. (2020), the authors use several information-theoretic notions that are central to our definitions. For example, they define individuality as the propagation of information from the past to the future and require low environmental determination as a characteristic of autonomous individuals. Krakauer et al. (2020) demonstrate how to apply entropy, mutual information, and partial information decomposition to analyze the system-environment interaction and capture different types of individuality.

Some limitations and future work

Although the analogy between, e.g., conflict level and entropy is appealing, there are a number of provisos attached to the possibility of applying an information-theoretic measure to human actions. Some were already noted by Berlyne (1957, p.332) such as, e.g., the objection that response tendencies might be simultaneous or that the alternative responses may not cancel each other completely, but rather their incompatibility can be seen as partially antagonistic in a way that one response does not exclude performance of another altogether. Berlyne himself offers some solutions to such problems. For now, we are satisfied with interpreting an action α in our model as α being selected, rather than executed. We are, of course, aware that complex agents can execute and do execute multiple simultaneous actions.

One of the major drawbacks of measures that rely on the agent-environment (or, *mutatis mutandis*, on the action-environment and system-environment) distinction is that the distinction already assumes knowledge of the interaction between the agent and the environment. (Bertschinger et al., 2008, p. 342) identifies the problem as follows: "The main difficulty, which exemplarily occurs in this case, is to define a suitable system–environment distinction with the corresponding observables. At this moment we are not able to provide a general method to perform this task, because this would require identifying the organization of the system (i.e. the autopoietic organization for an autopoietic system) algorithmically, which, at least implicitly,would solve the problem of a formal and operational definition of autopoiesis, which is not available by now." Our framework does not avoid the problem by shifting to the action-environment distinction. However, we are willing to concede that we do not have an answer to the problem at hand, as our goal is to use this distinction as heuristics in detecting the structure of intentions against the methodological 'scaffolding' of the action-environment distinction. Importantly, we do not claim to provide a general definition of autopoiesis, agency in an ontological sense, or the distinction between life and non-life.

Finally, an obvious criticism is that our framework lacks empirical validation. Although we have stated our goals here as primarily theoretical, we provide a blueprint for an information-theoretic intention detection algorithm. Specifically, this work lays the operational groundwork for applications using tools such as entropy estimators and solvers to determine whether an action is intentional or not and to what extent the output is reliable. Future work will operationalize the measures in controlled artificial and synthetic environments. Table 2 shows the expected qualitative output of comparing random, reactive, model-free, and model-based (reinforcement learning) agents in simple tasks. We predict

Agent	**conflict**	**control**	**memory**	**endogeny**	**complexity**
Random	*high*	*none*	*low*	*low*	*low*
Reflex	*low*	*none*	*low*	*low*	*low*
Model-free	*medium*	*some*	*medium*	*medium*	*medium*
Model-based	*medium*	*strong*	*high*	*high*	*high*

Table 2: Expected evaluation of (artificial) agents

monotonic increases in memory, (synergistic) complexity, and endogeny with increasing degrees of policy depth. These 'structural' dimensions will be combined with robustness tests using distracting events to approximate the system's resistance to misleading cues or its cognitive control. Finally, we pave the way for comparisons to canonical artificial life systems, such as cellular automata and evolving agents, in order to evaluate whether benchmark lifelike systems comply with the expected intentionality scores.

Conclusion

This paper proposes a five-dimensional definition of intentional action. Each dimension, namely, *conflict level*, *cognitive control*, *memory*, *endogeny*, and *pattern complexity*, has an associated information-theoretic measure. The uniqueness of our framework is that it offers a universal, structural account of intentions, independent of any agent-specific representation or internal model. The five-dimensional definition gives the required flexibility to analyze intentions from a purely behaviorist perspective as "current surrogates of the contingencies" (Skinner, 1984, p. 574) without oversimplifying the phenomenon of intention. Importantly, information-based measures provide a clear foundation for an intention-detection algorithm that will empirically validate the proposed account. Philosophically, we hold that detecting intentional action can be generalized across agent types only by asking 'how?', not 'why?'.

Acknowledgements

The authors are grateful for the valuable feedback from Rick Quax, Virgil Griffith, the attendees of the Chapman University Brain Institute meeting and the Utrecht University

GOALLAB colloquium, and the ALIFE 2023 program committee. This work was supported by the Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO) through the funding of the "Empowering Human Intentions through Artificial Intelligence" project awarded to Jan Broersen and Henk Aarts. The work of the first author is supported by the Italian Ministry of University and Research through the "Reasoning with Data" project (ReDa, G53C23000510001) awarded to Hykel Hosni under the FIS1 Advanced Grant scheme.

References

Alexander, W. H. and Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, 14(10):1338–1344.

Anscombe, G. E. M. (1963). *Intention*. Cornell University Press, 2nd edition.

Antusch, S., Custers, R., Marien, H., and Aarts, H. (2021). Intentional action and limitation of personal autonomy. do restrictions of action selection decrease the sense of agency? *Consciousness and Cognition*, 88:1–13.

Berlyne, D. E. (1957). Uncertainty and conflict: A point of contact between information-theory and behavior-theory concepts. *Psychological Review*, 64(6p1):329.

Bertschinger, N., Olbrich, E., Ay, N., and Jost, J. (2008). Autonomy: An information theoretic perspective. *Biosystems*, 91(2):331–345.

Bonchek-Dokow, E. and Kaminka, G. A. (2014). Towards computational models of intention detection and intention prediction. *Cognitive Systems Research*, 28:44–79.

Brass, M., Derrfuss, J., Forstmann, B., and von Cramon, D. Y. (2005). The role of the inferior frontal junction area in cognitive control. *Trends in Cognitive Sciences*, 9(7):314–316.

Brass, M. and Haggard, P. (2008). The what, when, whether model of intentional action. *The Neuroscientist*, 14(4):319–325.

Bratman, M. (1984). Two faces of intention. *The Philosophical Review*, 93(3):375–405.

Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79.

Combrisson, E., Basanisi, R., Neri, M., Auzias, G., Petri, G., Marinazzo, D., Panzeri, S., and Brovelli, A. (2025). Higher-order and distributed synergistic functional interactions encode information gain in goal-directed learning. *Nature Communications*, 16(1):7179.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Einstein, G. O. and McDaniel, M. A. (1990). Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4):717.

Fan, J. (2014). An information theory account of cognitive control. *Frontiers in Human Neuroscience*, 8:1–16.

Gardner, B., Arden, M. A., Brown, D., Eves, F. F., Green, J., Hamilton, K., Hankonen, N., Inauen, J., Keller, J., Kwasnicka, D., et al. (2023). Developing habit-based health behaviour change interventions: Twenty-one questions to guide future research. *Psychology & health*, 38(4):518–540.

Georgeff, M. and Rao, A. (1991). Modeling rational agents within a BDI-architecture. In Allen, J., Fikes, R., and Sandewall, E., editors, *Proceedings of the 2nd International Conference on Knowledge Representation and Reasoning (KR'91)*, pages 473–484. International Joint Conference on Artificial Intelligence (IJCAI Organization), Morgan Kaufmann.

Griffith, V. and Koch, C. (2014). Quantifying synergistic mutual information. In Prokopenko, M., editor, *Guided Self-Organization: Inception*, pages 159–190. Springer.

Hägele, A., Gema, A. P., Sleight, H., Perez, E., and Sohl-Dickstein, J. (2026). The hot mess of ai: How does misalignment scale with model intelligence and task complexity? *arXiv preprint arXiv:2601.23045*.

Han, T. A. (2013). *Intention Based Decision Making and Applications*, pages 55–74. Springer.

Herwig, A., Prinz, W., and Waszak, F. (2007). Two modes of sensorimotor integration in intention-based and stimulus-based actions. *Quarterly Journal of Experimental Psychology*, 60(11):1540–1554.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., and Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology*, 3(3):e79.

Ince, R. A. A. (2017). Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7):318.

Ji, J., Qiu, T., Chen, B., Zhou, J., Zhang, B., Hong, D., Lou, H., Wang, K., Duan, Y., He, Z., Vierling, L., Zhang, Z., Zeng, F., Dai, J., Pan, X., Xu, H., O'Gara, A., Ng, K., Tse, B., Fu, J., Mcaleer, S., Wang, Y., Yang, M., Liu, Y., Wang, Y., Zhu, S.-C., Guo, Y., Yang, Y., and Gao, W. (2025). AI alignment: A contemporary survey. *ACM Computing Surveys*, 58(5).

Kenton, Z., Kumar, R., Farquhar, S., Richens, J., MacDermott, M., and Everitt, T. (2022). Discovering agents.

Kolchinsky, A. and Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus*, 8(6):20180041.

Krakauer, D., Bertschinger, N., Olbrich, E., Flack, J. C., and Ay, N. (2020). The information theory of individuality. *Theory in Biosciences*, 139(2):209–223.

Krieghoff, V., Brass, M., Prinz, W., and Waszak, F. (2009). Dissociating what and when of intentional actions. *Frontiers in Human Neuroscience*, 3:1–10.

Mackie, M.-A., Van Dam, N. T., and Fan, J. (2013). Cognitive control and attentional functions. *Brain and Cognition*, 82(3):301–312.

Martela, F. (2025). Artificial intelligence and free will: Generative agents utilizing large language models have functional free will. *AI and Ethics*, 5(4):4389–4400.

McGregor, S. and Chrisley, R. (2020). The physical mandate for belief-goal psychology. *Minds and Machines*, 30(1):23–45.

McGregor, S., Virgo, N., et al. (2024). Formalising the intentional stance 1: Attributing goals and beliefs to stochastic processes. *arXiv preprint arXiv:2405.16490*.

Memelink, J. and Hommel, B. (2013). Intentional weighting: A basic principle in cognitive control. *Psychological Research*, 77(3):249–259.

Miller, E. K. (2000). The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience*, 1(1):59–65.

Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24(1):167–202.

Nachev, P., Rees, G., Parton, A., Kennard, C., and Husain, M. (2005). Volition and conflict in human medial frontal cortex. *Current Biology*, 15(2):122–128.

Orseau, L., McGill, S. M., and Legg, S. (2018). Agents and devices: A relative definition of agency.

Prokopenko, M., Boschetti, F., and Ryan, A. J. (2009). An information-theoretic primer on complexity, self-organization, and emergence. *Complexity*, 15(1):11–28.

Quax, R., Har-Shemesh, O., and Sloot, P. M. (2017). Quantifying synergistic information using intermediate stochastic variables. *Entropy*, 19(2):1–27.

Rosas, F. E., Mediano, P. A. M., Rassouli, B., and Barrett, A. B. (2020). An operational information decomposition via synergistic disclosure. *Journal of Physics A: Mathematical and Theoretical*, 53(48):1–27.

Searle, J. R. (1980). The intentionality of intention and action. *Cognitive Science*, 4(1):47–70.

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

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Skinner, B. F. (1984). Coming to terms with private events. *Behavioral and Brain Sciences*, 7(4):572–581.

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7):450–461.

van Zee, M., Doder, D., van der Torre, L., Dastani, M., Icard, T., and Pacuit, E. (2020). Intention as commitment toward time. *Artificial Intelligence*, 283:103270.

Varley, T. F. and Hoel, E. (2022). Emergence as the conversion of information: A unifying theory. *Philosophical Transactions of the Royal Society A*, 380(2227):20210150.

Wibral, M., Lizier, J. T., and Priesemann, V. (2014). Bits from biology for computational intelligence. *arXiv preprint arXiv:1412.0291*.

Williams, P. L. and Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Xiong, W., Faes, L., and Ivanov, P. C. (2017). Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations. *Physical Review E.*, 95(6):062114.

Zhang, Z., Zeng, Y., Jiang, W., Pan, Y., and Tang, J. (2023). Intention recognition for multiple agents. *Information Sciences*, pages 360–376.