ACTIONABLE ASPECTS OF AGENCY

Izak TAIT

ABSTRACT: This paper proposes a comprehensive classification scheme for identifying agentic entities based on six essential aspects: perception, predictive capacity, decision-making capability, goal-content integrity, resource input and output, and manipulators. By systematically examining these aspects, this paper aims to distinguish true agents from non-agentic entities, including advanced tools and collective entities. The classification scheme addresses practical implications, such as legal liability and moral considerations for AI models and the potential identification of extraterrestrial life forms. Perception is identified as the foundational aspect, enabling the acquisition and processing of environmental information. Predictive capacity and decision-making capability allow entities to anticipate outcomes and select optimal actions, respectively. Goal-content integrity ensures sustained commitment to intended goals, while resource input and output highlight the necessity of resource management. Manipulators are essential for the physical realisation of intentional actions. The paper emphasises the interconnectedness of these aspects, demonstrating that their integration is required for an entity to be classified as an agent. This framework provides valuable insights for diverse fields, from legal and ethical discussions surrounding AI to philosophical debates on personhood and agency, enhancing our understanding of intelligent entities across various contexts.

KEYWORDS: agency, intentionality, action, agent

1. Introduction

An entity can only be classified as an agent if it meets the requisite criteria of an agent. Tautological as this is, it does imply that there are certain attributes and characteristics that separate an agentic entity from a non-agentic entity. This paper will detail six of these attributes, named 'aspects', that are required (and together, potentially sufficient) for an entity to be classified as an agent.

This paper aims to create an actionable classification scheme that may be used to categorise tools (even 'smart' tools) from true agentic entities. While all humans are (generally speaking) classified as agents, a classification guide is essential in determining whether non-human, non-vertebrate entities are agents. In cases of collective entities, such as companies and corporate entities, being classified as agentic has legal consequences (Duff 1990); however, for invertebrate animals and digital entities, such as AI, there may be a moral component to the classification of an agent as it implies a degree of reciprocity with its environment (Parthemore and Whitby 2013).

In the particular case of AI entities, the case for legal liability and protection may hinge on whether a specific model is classified as simply a tool or as an agent in its own right (Dai 2024). More speculatively yet, should humanity ever encounter complex extraterrestrial life, there will be no shared biological and cultural evolutionary history to compare the extraterrestrial entities with terrestrial agents. This paper's classification scheme would thus be able to serve to determine whether what we may encounter is an agent or just a tool.

On the philosophical side, agency is a core component of personhood (Pietrzykowski 2018; Dennett 1988; Taylor 1985). Thus, should an entity display all the attributes and characteristics of agency discussed below, they have reached a milestone towards being considered a person.

This paper will outline the requisite aspects of agency by focusing on what is required to perform an intentional action. Through this specific perspective, this paper will avoid commenting on the debate between free will, compatibilism, and determinism. The paper will also not presume to take a side between the event-causal, the agent-causal, and the volitionist frameworks of agency. For details on these, we recommend Schlosser's excellent article on the topic of agency (Schlosser 2019). Whatever the reader's views may be on the metaphysical nature of agency, the aspects of agency laid out below ought to be compatible.

Lastly, the paper will avoid the issue of deviant causal chains by focusing on what is required to perform immediate basic actions, rather than processes of actions in what may be referred to as non-basic actions. The aspects below are applicable to long-term actions (as running examples will show); however, the intentionality of the performance, rather than the execution, of the actions will be the focus of this paper.

2. Aspects of Agency

Each of the subsections below will focus on a specific aspect of agency (summarised in Table 1 through lay-descriptive and narrative examples for ease of understanding) and discuss why each is required for an entity to be considered an agent capable of performing an intentional action, yet why each aspect is not sufficient by itself.

Because of this, it should always be borne in mind that there will be many non-agent entities that have one, or several, of the aspects below. However, if an entity does not at least have all six aspects below, it cannot be classified as having the capacity to perform intentional actions and, thus, agency.

Note that the aspects below are not ordered in any sense of importance, significance or hierarchy. Much like the interlocking connecting construction toys, the aspects of agency may be ordered and arranged in whatever formation an entity requires to express its agency.

Table 1: Attributes and characteristics that are required for agency.

Attributes	Description and Example
Perception	The ability to receive and process information in the environment. "I see an apple."
Predictive Capacity	Anticipating the probabilities of success of a set of actions or goals. "There are several ways I can obtain the apple."
Decision-Making Capability	Evaluating available options and selecting a course of action or goal. "I know how to obtain the apple."
Goal-Content Integrity	Maintaining commitment to and coherency of a goal throughout the performance of an action. "I focus on the apple while reaching for it."
Resource Input and Output	Receiving and expending resources to perform an action. "I expend energy to reach the apple."
Manipulators	The means through which an entity performs an action. "My hand reaches for the apple."

Before continuing, it is essential to define the terms that will be used throughout the paper:

- A goal is defined as a specific and measurable state or condition that serves as an endpoint or target, resulting in an altered state of the environment when achieved.
- An action is defined as any change made by an entity in the state of the environment (or objects/entities therein) over time.
- Thus, an intentional action is an act performed with a specific goal or purpose

in mind, guided by an agent's desires, beliefs, and intentions.

- If an entity can perform intentional actions, it has agency and can be classified as an agent.
- An agent is thus a subset of the set of all entities that can perform actions.

2.1. Perception

1. Performing an intentional action requires interaction with the environment.
2. Interacting with the environment requires the coordination of an entity's outputs within the environment.
3. Output coordination requires processing information about the environment.
4. Processing information about the environment requires perceiving information in the environment.
5. Performing an intentional action requires formulating a goal to change the environment.
6. Formulating a goal to change the environment requires acquiring information on the current and desired state of the environment.
7. Acquiring information about the environment requires perception.
8. Ergo, perception is required for agency.

Perception is the cornerstone of agency. As crucial as all the aspects below are to performing intentional actions, the process of agency begins with the perception of the environment. If an entity is unable to perceive its environment (whether its external, internal or mental environment), then it will be unable to perform an intentional action. This is because of two reasons.

Firstly, an intentional action affects a change in the environment based on a goal to do the same. Without perception, an entity is unable to acquire and interpret information on the current state of the environment to formulate a goal to change it. This links perception directly with the aspects of predictive capacity and decision-making ability, as these are involved in the formation of intentional goals.

Therefore, perception is necessary for there to be intentionality and goal formation.

Note that unintentional actions do not require perception as a rule, as these are not (necessarily) goal-driven. A printer does not need to perceive where the paper is or where its ink cartridges are to print; it will always print within the same margins. Equally, perception does not necessitate an action (intentional or not). A security camera perceives its environment through its video and audio recording, but it does not by itself perform actions.

Additionally, the entity’s perception does not need to be phenomenally conscious in the sense that it gives rise to mental states with qualitative, subjective states with phenomenal content. The perceptive computation may operate entirely on a mechanistic or functional level, processing information and relating the necessary actions to the goal at hand, without any valent feelings or phenomenal value placed on what is being perceived.

The second reason for perception's inclusion in this list is that perception is necessary for the continuous feedback the agent needs about the environment to adjust the intended action in real-time to ensure the goal is achieved (Bertenthal 1996; Creem-Regehr and Kunz 2010; Halász and Cunnington 2012). This continuous feedback loop is maintained through the entity’s perception of itself and its environment (Hurley 2001; Hayhoe et al. 2020). This feedback is not solely about the physical movements of the entity, but also about its intended goal, linking perception to the aspects of goal-content integrity and the entity’s manipulators.

The perceptual feedback links the agent and its environment together into a dynamic information system (Warren 2006, 1990), whereby the agent changes the state of the environment, generating new information which is perceived by the agent to alter or modulate its movements and actions.

This feedback does not extend to an entity needing to perceive the success or failure of its goal in order to be classified as having this aspect of agency. An entity merely needs to begin its intentional action towards its goal. For example, shooting someone with a firearm is an intentional action; however, pulling the trigger and then being prevented from perceiving the outcome would not negate the intentionality or the agency of the entity.

To see the link between perception and intentional actions, we can begin with the examples that will be used throughout this paper.

For the natural agent, Bob perceives his thirst via interoception, the cup of tea through exteroception and his arm through proprioception. With all of this, he can formulate the goal of picking up the cup of tea to drink it, quenching his thirst. As Bob moves his arm, the proprioception of his arm and exteroception of the environment allow him to adjust the movements of his arm as it moves to the cup and brings it back to his mouth, completing the action.

For the artificial entity, an LLM perceives the prompt from the user, initiating its action of completing its response. It perceives the changes to the input data as it passes through internal weights when processing a response, and perceives the response itself as it is processed. An LLM is far more constrained in its capacity to modulate its actions, compared to a collective or natural entity; however, perception of its limited environment is equally crucial.

For an intentional action on a more extended timeframe that requires several intermediary actions, we can look at an ant colony defending itself from an attacker. The individual ants that comprise this collective entity would perceive attackers, and communicate this through the colony such that there is a collective sense of awareness of a hostile intruder. A response would similarly filter its way through the colony to the soldier ants, which would initiate a defence of the colony by attacking the intruders. The perceptual feedback loop of the action would be, again, the individual ants perceiving the current state of the 'battle' and communicating this to the rest of the colony.

Perception is, thus, necessary for agency as it allows an entity to gather information, form goals, and adjust actions based on feedback. This leads to predictive capacity, where achieving a desired result requires inferring effective behaviours and considering probabilities based on perceptive input and internal information.

2.2. Predictive capacity

1. Performing an intentional action requires aiming for a desired result.
2. Achieving a desired result requires inferring the most effective behaviour to reach that result.
3. This inference requires (pre)reflective consideration of the probability of success.
4. Considering the probability of success is based on perceptive input and internal information.
5. This consideration allows an entity to predict the outcome of its behaviours in relation to the desired result.
6. Ergo, a predictive capacity is required for agency.

For every possible goal and intention, there necessarily exists a large (but practically limited) set of methods or processes that can be attempted to meet that goal successfully. For each possible method, there exists a set of actions that can potentially be used. As the idiom says, "There is more than one way to skin a cat."

An entity cannot simultaneously attempt every possible action for every possible method to achieve a goal. For example, Bob cannot be both standing up to reach for his cup of tea, while also remaining sitting to do the same; he cannot use his right hand to reach for the cup while also keeping his right hand on the chair to help him stand up. An entity must (pre)reflectively and prospectively anticipate and infer the possible outcomes of any proposed action and method and select the one most likely to achieve the desired goal (Bertenthal 1996). Thus, areas of Bob's brain, such as the visual cortex, motor cortex and cerebellum, work together to predict the optimal movements that Bob's muscles need to perform in order to pick up the cup of tea and bring it to his mouth. Bob, in all likelihood, would not be conscious of these computations as these would be automatic to Bob's natural movements.

The predictive capacity does not need to extend to the end of the intended goal, merely to the most probable action to meet the goal's criteria. As an example, an ant colony mounting a defence against an intruder does not need to have the collective capacity to accurately predict whether its intended actions would successfully repel the attacker. Equally, Bob does not need to have the clairvoyance of whether his intended motion will be able to reach the cup. The predictive capacity only needs to extend to what is most probable. Bob may have perceived the cup as being closer than it truly was, resulting in him not reaching far enough to pick it up. The ant colony may have only collectively perceived a small number of attackers, thus sending a small number of soldier ants to respond, not knowing that their intruders were numbered by the legion.

While the above is enough to satisfy the minimum requirements for this aspect of agency, in human society (particularly in civil and criminal law), the capacity to predict the outcomes of an action is crucial to assigning accountability to any intentional action (W. Chan and Simester 2011; Yaffe 2004). Knowing the consequences of an action makes an agent accountable for that action. The further in time, and the higher order, an agent is able to predict its action's consequences, the higher it may be held liable for those actions. This is, in part, why minors, those with mental deficiencies, and animals are held to a lower standard of accountability when it comes to assigning liability for their actions (Höglund et al. 2009; Loomis-Gustafson 2017).

As with all the aspects of agency, the capacity to predict future states or outcomes based on current knowledge and information does not equate to agency by itself. Weather forecasting models can accurately predict future weather events by processing historical weather data and current atmospheric conditions to output a forecast that is (at times) reliable. This does not mean, however, that current weather forecasting software and systems are agents.

On the other hand, other software and machine models may be classified as agents. LLMs also use prediction to select the appropriate response to input prompts. LLMs generate coherent and contextually relevant text by anticipating the most probable next word or phrase based on extensive training data. This predictive mechanism is integral to their functioning, and the reason for their commercial success.

Predictive capacity is required not only for performing an action but also for the formation of goals themselves. All goals are a target or end-point to meet a desired change in the environment. Any desire may have more than one possible goal that could meet its targetted change in the environment. The agent's predictive capacity allows it to anticipate which goal may be the most profitable.

A predictive capacity, by definition, requires an agent to have a memory and learning capacity of some sort (Estevez and Calvo 2000; Barron, Auksztulewicz, and Friston 2020; Shing, Brod, and Greve 2023). Predicting the most profitable course of action requires knowledge of the agent and the environment, in turn requiring that the agent must have learnt something of both. An agent in a completely unknown environment would not have any knowledge of the environment, but would still have knowledge about itself which it could use to predict what it can and cannot do (Hayhoe et al. 2020).

The knowledge required to predict would have had to come from learning; however, as current LLMs show, this learning and memory may be accomplished through training data and be entirely preloaded into the entity. Additionally, the extent to which an agent can anticipate future events will dictate how much memory it needs and the learning it is required to have undergone. As such, for a minimal sense of predictive capacity, an agent may only need to have a memory of itself and have learnt (actively, passively, front-loaded or not) how to use its own manipulators to accomplish goals.

Performing an intentional action requires aiming for a desired result and predicting the most effective behaviour to achieve it. This predictive capacity is crucial for both action execution and goal formation, allowing an entity to anticipate outcomes and select the most probable method. Consequently, the decision-making capability is inherently linked to predictive capacity, ensuring the agent commits to the most suitable action to achieve its goals.

2.3. Decision-making capability

1. Performing an intentional action requires matching a behaviour to a desired result.
2. Matching a behaviour to a desired result requires selecting one behaviour from multiple possible options.
3. Selecting one option from multiple alternatives requires evaluating and comparing those alternatives.
4. Evaluating and comparing alternatives requires a decision-making process.
5. Ergo, a decision-making capability is required for agency.

The capability to make decisions is intimately tied to the capacity to predict outcomes in that both lead to the same result: selecting the action with the most probable perceived chance of success. How these two differ is in the high-level foundational processes themselves. An entity may easily have a predictive capacity without the capability to make decisions and vice versa.

As mentioned above, predictive capacity allows an entity to anticipate future states or outcomes based on current knowledge and information. It can generate what the most seemingly correct decision would be, but this does not mean it has the capacity to select that decision. Selecting a specific course of action, whether the one most profitable or not, is what allows an agent to commit to that course of action and physically, digitally, or mentally perform the action.

This aspect is crucial not only for any intentional action in question but also for goal formation itself (Harter 2006; Bagozzi 1997). A goal is the desired change to the environment to meet a need that the current environmental state lacks. Bob's thirst is an undesirable environment state which he desires to change. His goal, therefore, is to pick up the cup of tea and drink it, thereby altering the environmental state to meet the initial need.

However, as with the previous section, for any desired state, there is a large (but limited) set of possible goals that may or may not meet that state's criteria. Like deciding on a specific action, predicting which goal will meet that desired state belongs to the previous aspect, but deciding on that goal is core to this aspect.

This decision-making capability is fundamental to the concept of agency, as it underscores the entity's ability to be the driver of its own goals and actions. In humans, this is often called volition and is vital in determining whether an entity is in control of its actions (Dijksterhuis and Aarts 2010; Liljenström 2022; Tait 2024). Should a secondary entity be determined to have created both the goal and the action for the primary entity, then the primary entity cannot be said to be in control of its actions; the secondary entity would be in control. If an entity does not have the capability to reject or refuse a goal and its associated set of actions, then it does not have the capability to make a decision.

For example, modern LLMs, via their chat interfaces (such as ChatGPT), cannot craft goals nor decide upon their actions. Their goal is always to respond to a user's prompt using the most probable response that fits within the creator's moderation guidelines for acceptable speech. The actions that the LLMs take in this environment are equally limited. The same process is repeated for each prompt, with the next word selected via an algorithm based on its probability value. LLMs cannot refuse or change their goals or actions. One can, therefore, say that LLMs do not have a decision-making capability in their native configuration.

And yet, LLMs have been used as agents by utilising second-order effects. LLMs have been used to play video games and operate robots by giving them access to external tools and the space in which to operate these via their next-token-predicted stream of text (de Wynter 2024; Hu et al. 2024; Lifshitz et al. 2023). Thus, while they cannot change their own goals and actions, by providing them with a scenario and, therefore, a vehicle to roleplay, they can be shown to be agentic in their interactions with other tools. Input from these tools (robots, video games, etc.) would further prompt the LLM, which cannot refuse or decide how to respond, but its output response can include decision-making as to the external tool. These second-order agentic effects can thus be used to classify an LLM as having decision-making capabilities.

Similarly, in collective entities such as an ant colony, each individual ant is not making the decision to respond to an attack on the colony, or which action to take to respond appropriately. Instead, it is the communications between each individual ant and the network of the colony that produces a collective decision-making capability to respond to environmental stimuli (Bose, Reina, and Marshall 2017; Marshall et al. 2009; Edwards and Pratt 2009). The colony responds even if each individual member cannot cognitively process the decisions.

In contrast, an example of a non-agent decision-making entity would be an automated thermostat. Tasked with a singular goal of maintaining a room's temperature, the thermostat 'decides' whether to turn on the heating or the cooling based on the environmental stimuli. At each point in time, the thermostat has the capability to choose between heating and cooling (and the degree to which it does either), and based on its perception, makes the decision as to which.

Selecting a specific course of action allows an agent to commit to and perform that action, making decision-making fundamental to agency. This capability is essential for maintaining goal-content integrity, as the next section will show, which requires a sustained commitment to a goal throughout its execution, preventing distractions. Together, decision-making and goal-content integrity enable an entity to perform intentional actions and maintain agency.

2.4. Goal-content integrity

1. Performing an intentional action is done in service to a goal.
2. Performing an intentional action requires a unit of time to elapse.
3. This unit of time spans from when an entity decides on a goal until the entity's manipulation of the environment is complete.
4. During this unit of time, the entity must maintain its commitment to the initial goal to complete the action.

5. Valuing another goal and shifting attention or resources would hinder the completion of the present action.

6. Ergo, goal-content integrity is required for agency.

An intentional action implies a directed change to the environment towards a desired state, meaning that every intentional action is intrinsically linked to a goal as that goal is the desired state (Synofzik, Vosgerau, and Newen 2008; Bello and Bridewell 2020). Goals are crucial not only for the formation of an intentional action but also for the agent's commitment to perform that action.

If an entity's immediate goal changes in the midst of committing to a goal, then its intention would likewise change. A change of intention would mean its current action is no longer in service to a desired changed state in the environment. All of this would hinder the completion of the action, and could prevent it altogether. For example, should Bob get distracted and his attention be captured by a can of soda, his motivation and goal may change to wanting that instead. Thus, he may stop his action of picking up the cup of tea, and move to pick up the can instead.

A longer-term example would be of the ant colony responding to an attacker. Should the network of pheromonal communication within the colony become disturbed by signals from ants not involved in the colony's defence, the colony's emergent attention may shift towards the content of the novel communications, hindering its defence as its pheromonal communication directed towards it loses coherence.

Attention is a key driving force in goal formation, goal-content integrity, and goal-switching. Sustained attention is crucial for ensuring consistency and coherency of the actions and content of the goal (Dijksterhuis and Aarts 2010). Attention enables agents to detect discrepancies or errors in the goal-related processes, allowing for timely adjustments and alignment of the action with the goal's original content and purpose (Diehl, Semegon, and Schwarzer 2006; Luszczynska et al. 2004).

Diverting attention away from the current task and goal, intentionally or otherwise, can lead to a shift in priorities and motivation to what is newly being attended to. This can result in the original goal being neglected or downgraded in importance, affecting its overall integrity and the likelihood of the action's completion. As the action no longer supports the (newly) immediate goal, resources and energy may no longer be spent on it.

Note that in order for an entity to attend to a subject within its environment, it requires the ability to perceive. This means that entities without the capacity for attention may not be able to have their attention shifted. Thus, such an entity would logically seem always to keep its goal-content integrity secure. An example of this would be the pendulum clock, which maintains its goal of keeping time accurately via its mechanism that continuously oscillates and regulates its movement to ensure consistent timekeeping.

LLMs may be another such example. As noted above, in their native configurations through their chat-interfaces, an LLM is locked into a goal of responding to user input with the most probable sequence of text that follows. As such, while transformer-based LLMs use attention extensively in their computation of their response text, once they have begun computing a response, they will continue with their response unless there is a fault in the system or the user stops the process. However, as a conversation continues (particularly beyond an LLM's context window), the AI model may begin to hallucinate more frequently and lose the initial purpose and goal of the conversation. Therefore, through its hallucinations and exceeding the context window in which it can accurately read text, it may lose the capacity to secure its goal-content integrity.

Coupled with attention is the need for an entity to have a working memory in order to ensure goal-content integrity. For an agent to attend to its goal over any period of time, it requires a cognitive unit or system that can maintain and process transient information (Estevez and Calvo 2000; McVay and Kane 2009). If such a cognitive unit did not exist, the agent would be unable to maintain the perceptive information required for attention. An agent, therefore, must, at the very least, have a working memory sufficient enough to maintain information for the duration of the action's execution.

While one may argue that working memory is required for a minimal sense of agency, if the memory capacity is required to equal the length of time to execute the action, this means that long-term actions will require long-term memory. Thus, one can place memory capacity on a spectrum, with working memory on one end for the minimal sense of agency through goal-content integrity, and long-term memory on the opposite side of the spectrum, required for complex agents performing complex actions and goals.

Ensuring an agent's goal-content integrity remains stable is not merely a passive matter but requires active and ongoing consideration. If an agent is prevented from completing its action by an outside force, then its goal is not achieved. If Alice takes Bob's cup of tea before he can reach it, then he cannot drink it to quench his thirst. Therefore, an agent is incentivised to prevent other entities or events from interfering with its actions (Bostrom 2012; Hurley 2001; Omohundro 2008). This is more crucial for long-term actions such as the ant colony's defence (which may have intermediate steps) rather than short-term immediate goals such as Bob and his tea (as Bob's attempt to prevent Alice from getting his cup is a separate goal and associated set of actions which will require him to stop his current task). Paradoxically, this highlights how active goal-content integrity may work against itself in certain situations.

Goal-content integrity requires that an agent maintains a consistent focus on its desired outcome throughout the action, preventing shifts in attention that could disrupt goal completion. Similarly, the next section shows how performing an intentional action requires the acquisition and expenditure of resources, such as energy, to sustain the efforts needed to achieve a goal.

2.5. Resource Input and Output

1. Performing an intentional action requires the expenditure of resources, such as energy.
2. Resource expenditure requires a system through which resources can be moved towards necessary outputs.
3. A resource-based system requires the input of resources to function.
4. Ergo, resource acquisition and expenditure are required for agency.

This aspect deals more with the fundamental laws of the physical universe than it does with the concepts of agency; however, its consequences are as crucial. An action, intentional or not, is the use of energy to affect a change in the environment. We can see this by gradually, and more granularly, defining the terms at hand.

A change in the environment via an action requires work, defined by physicists as the product of force and displacement. Force is necessary to change the state of motion of an object, which, according to Newton's first law of motion, is uniform unless acted upon by an external force. Interaction by any two bodies (or entities) transfers energy from one body to the other (to affect its motion), or from one form to another (e.g., kinetic energy, potential energy, thermal energy). As the first law of thermodynamics prohibits the creation (or destruction) of energy, this energy transfer is a necessity rather than a convenience.

Therefore, an action (intentional or not) cannot occur without the transfer or transformation of energy. Yet, more than this, for an agent to perform an action, that agent requires the energy for it to be transferred or transformed in the action. Bob can only pick up his cup of tea by expending the necessary joules of chemical energy to move the muscles in his body to affect that action. An LLM equally cannot respond to user inputs without electricity powering its physical computer hardware. Even non-agents require this, as a Wisteria requires energy to climb and spread its vines, which it fortunately acquires from the sun via photosynthesis.

This last point also shows the other side of this aspect. For an agent to expend energy, it must also be able to acquire energy, else it would break the first law of thermodynamics.

However, it is not only energy which is the sole resource an agent requires, even if it is generally perceived as the predominant one. Other resources exist that, in specific contexts and types of agents, would prohibit an action from being performed in their absence. An LLM cannot respond to its users' inputs without the vast quantities of data in its datasets. Data, to an LLM (and other machine agents), are a crucial resource that it requires to carry out its action; and macro and micronutrients are as vital to Bob being able to use his muscles as the energy they consume to move.

In the same vein, to a collective agent, such as the ant colony, each of its constituent members is a keystone resource. While each ant acquires and expends energy in its action, the colony itself requires the acquisition (through breeding) and expenditure of ants to carry out its collective actions.

Resources need not be physical for them to be necessary for an action. Knowledge and power are often overlooked as key to most mid to long-term actions. If Bob discovers his tea requires sugar, he needs to know where the sugar is to obtain it. To defend itself from an attacking force, the ant colony requires physical and numerical power to overcome its intruders. In AI, power-seeking behaviour has been speculated upon greatly as a fear of how superintelligent AI may be an existential risk to humanity (Bostrom 2012; Omohundro 2008) and that AI's perception of the 'necessary' level of power may not align with humanity's, as it may be incentivised through its reward mechanisms to seek power (Turner et al. 2019; A. Chan et al. 2023).

Other than the necessity of energy for all actions, the resource required for any action is contextual; however, the entity is required to be able to acquire and expend it to be classified as an agent. Resource acquisition would also, therefore, become a recurring goal and subject of an agent's intentional actions (in the short and long term).

Resource acquisition and expenditure are essential for an agent to perform intentional actions, requiring energy and other resources to sustain efforts towards a goal. This necessity of interacting with the environment at a fundamental level parallels the need for manipulators, as an agent must interact with its environment to execute these actions.

2.6. Manipulators

1. Performing an intentional action requires interaction with the environment.

2. Interacting with the environment requires a means through which an entity can manipulate the environment.
3. Ergo, manipulators are required for agency.

This aspect is at the opposite end of the agency spectrum from Perception, and so has been appropriately placed at the opposite end of this list of aspects. Where perception begins the cascade of internal processes and procedures that leads to agency, an agent's manipulators are the final section of that cascade; they are the part of the agent which physically performs the intentional action.

No matter what action an agent intends to perform, it must have a means through which it can perform that action. There must be something in the universe under the agent's direct control (nearly always a part of the entity itself) which can (attempt to) perform the intentional action that the agent believes will achieve its goal. Any interaction that the agent intends to have with the universe (whether inside its embodiment or in its external environment) requires a means and method with which to carry out that interaction. As noted in Section 2.5 above, any manipulation of the environment either necessitates an application of force or a transformation of energy. The means and method through which the entity does this is its manipulators.

This does not need to be a physical actuator such as a limb, although Bob does benefit from his own actuator (his arm), which can extend (supported by his shoulders and trunk) to reach out and grasp his cup of tea. It could be the neurons in a person's brain firing to create a thought or imaginative mental scenario, or the software in an LLM with which it can interface with a chat-client to interact with a user and respond to user inputs. A manipulator could also be on the more macro scale, such as each ant in an ant colony being a manipulator of the colony's emergent agency.

Manipulators come in all shapes and sizes, and a single agent may have many different manipulators to account for a variety of different tasks. Any (external) part of a human body may be used as a manipulator given the proper context, and an ant colony has as many manipulators as it has ants. Multiple manipulators may be used to achieve the same goal (Bob can use either of his hands to pick up his cup of tea, or attempt a more acrobatic feat by using his feet), and goals are often created to align with an agent's manipulators (Bob cannot fly, and thus would not reasonably create an intentional goal to fly).

Due to its ubiquity in the performance of actions, manipulators are seen in nearly all non-agent actors, those entities which act upon the environment without agency. This may be microscopic unicellular organisms interacting with their fluid environment through cilia and flagella, factory single-purpose robotic armatures performing the same action indefinitely, or carnivorous plants whose traps are activated by their prey.

3. Conclusion

The classification of an entity as an agent is contingent upon its possession of the six outlined aspects: perception, predictive capacity, decision-making capability, goal-content integrity, resource input-and-output, and its manipulators. Each of these aspects is individually necessary and collectively potentially sufficient to fulfil the criteria for agency, should they be successfully integration. The practical implications of this classification scheme extend to diverse domains, from legal frameworks determining liability and protection for AI models to the philosophical considerations of personhood and moral agency in non-human entities.

Perception serves as the foundation of agency, enabling entities to acquire and process environmental information necessary for intentional action. Predictive capacity allows entities to anticipate outcomes and select the most effective behaviours to achieve their goals. Decision-making capability facilitates the evaluation and selection of actions, ensuring that entities can commit to a course of action. Goal-content integrity ensures sustained commitment to an intended goal, while resource input and output highlight the necessity of acquiring and expending resources to perform actions. Finally, manipulators enable entities to interact with their environment, effectuating the physical realisation of their intentions.

The integration of these aspects provides a robust framework for distinguishing agentic entities from non-agentic ones. This framework has significant implications for our understanding of agency in various contexts, including artificial intelligence, collective entities, and potential extraterrestrial life. By applying this classification scheme, we can better navigate the ethical, legal, and philosophical challenges posed by the evolving landscape of intelligent entities.

Future research should focus on empirically validating this classification scheme across different types of entities, from advanced AI systems to complex biological organisms. Additionally, exploring the nuances of how these aspects interact in varied contexts will provide deeper insights into the nature of agency. Practical applications of this framework could include developing more sophisticated AI systems capable of exhibiting true agency, as well as refining legal and ethical guidelines for the treatment and responsibilities of non-human agents.

Ultimately, this paper presents a comprehensive and actionable classification scheme that enhances our ability to identify and understand agency across a spectrum of entities. This contributes to ongoing discussions in both theoretical and applied domains, fostering a deeper appreciation of the complex nature of agency and its manifestations in the world around us. By integrating these insights, we can advance our understanding and management of the dynamic relationships between human and non-human agents in an increasingly interconnected world.

References

Bagozzi, Richard P. 1997. "Goal-Directed Behaviors in Marketing: Cognitive and Emotional Perspectives." *Psychology & Marketing* 14(6): 539-43.

Barron, Helen C., Ryszard Auksztulewicz, and Karl Friston. 2020. "Prediction and Memory: A Predictive Coding Account." *Progress in Neurobiology* 192(101821): 101821.

Bello, Paul, and Will Bridewell. 2020. "Attention and Consciousness in Intentional Action: Steps Toward Rich Artificial Agency." *Journal of Artificial Intelligence and Consciousness* 07(01): 15-24.

Bertenthal, B. I. 1996. "Origins and Early Development of Perception, Action, and Representation." *Annual Review of Psychology* 47(1): 431-59.

Bose, Thomas, Andreagiovanni Reina, and James A. R. Marshall. 2017. "Collective Decision-Making." *Current Opinion in Behavioral Sciences* 16(August): 30-34.

Bostrom, Nick. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* 22(2): 71-85.

Chan, Alan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, et al. 2023. "Harms from Increasingly Agentic Algorithmic Systems." *arXiv [cs.CY].* arXiv. http://arxiv.org/abs/2302.10329.

Chan, Winnie, and A. P. Simester. 2011. "FOUR FUNCTIONS OF MENS REA." *The Cambridge Law Journal* 70(2): 381-96.

Creem-Regehr, Sarah H., and Benjamin R. Kunz. 2010. "Perception and Action." *Wiley Interdisciplinary Reviews. Cognitive Science* 1(6): 800-810.

Dai, Jessica. 2024. "Position: Beyond Personhood: Agency, Accountability, and the Limits of Anthropomorphic Ethical Analysis." In *Proceedings of the 41st International Conference on Machine Learning* 235: 9834-9845.

Dennett, Daniel. 1988. "Conditions of Personhood." In *What Is a Person?,* edited by Michael F. Goodman, 145-67. Totowa, NJ: Humana Press.

Diehl, Manfred, Angelenia B. Semegon, and Ralf Schwarzer. 2006. "Assessing Attention Control in Goal Pursuit: A Component of Dispositional Self-Regulation." *Journal of Personality Assessment* 86(3): 306-17.

Dijksterhuis, Ap, and Henk Aarts. 2010. "Goals, Attention, and (un)consciousness." *Annual Review of Psychology* 61: 467-90.

Duff, Robin Anthony. 1990. *Intention, Agency and Criminal Liability: Philosophy of Action and the Criminal Law*. Blackwell.

Edwards, Susan C., and Stephen C. Pratt. 2009. “Rationality in Collective Decision-Making by Ant Colonies.” *Proceedings of the Royal Society B: Biological Sciences* 276(1673): 3655-61.

Estevez, A., and M. G. Calvo. 2000. “Working Memory Capacity and Time Course of Predictive Inferences.” *Memory* 8(1): 51-61.

Halász, Veronika, and Ross Cunnington. 2012. “Unconscious Effects of Action on Perception.” *Brain Sciences* 2(2): 130-46.

Harter, D. 2006. “Complex Systems Approaches to Emergent Goal Formation in Cognitive Agents.” In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, 4966–71. IEEE.

Hayhoe, Mary, Katja Fiehler, Miriam Spering, Eli Brenner, and Karl R. Gegenfurtner. 2020. “Introduction to Special Issue on ‘Prediction in Perception and Action.’” *Journal of Vision* 20(2): 8.

Höglund, Pontus, Sten Levander, Henrik Anckarsäter, and Susanna Radovic. 2009. “Accountability and Psychiatric Disorders: How Do Forensic Psychiatric Professionals Think?” *International Journal of Law and Psychiatry* 32(6): 355-61.

Hurley, Susan. 2001. “Perception And Action: Alternative Views.” *Synthese* 129(1): 3-40.

Hu, Sihao, Tiansheng Huang, Fatih Ilhan, Selim Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. 2024. “A Survey on Large Language Model-Based Game Agents.” *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2404.02039.

Lifshitz, Shalev, Keiran Paster, Harris Chan, Jimmy Ba, and Sheila A. McIlraith. 2023. “STEVE-1: A Generative Model for Text-to-Behavior in Minecraft.” Edited by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. *Neural Information Processing Systems* abs/2306.00937 (June): 69900-929.

Liljenström, Hans. 2022. “Consciousness, Decision Making, and Volition: Freedom beyond Chance and Necessity.” *Theory in Biosciences* 141(2): 125-40.

Loomis-Gustafson, Carly. 2017. “Adjusting the Bright-Line Age of Accountability within the Criminal Justice System: Raising the Age of Majority to Age 21 Based on the Conclusions of Scientific Studies Regarding Neurological Development and Culpability of Young-Adult Offenders.” *Duquesne Law Review* 55: 221.

Luszczynska, Aleksandra, Manfred Diehl, Benicio Gutiérrez-Doña, Patrik Kuusinen, and Ralf Schwarzer. 2004. “Measuring One Component of Dispositional Self-

Regulation: Attention Control in Goal Pursuit*.” Personality and Individual Differences* 37(3): 555-66.

Marshall, James A. R., Rafal Bogacz, Anna Dornhaus, Robert Planqué, Tim Kovacs, and Nigel R. Franks. 2009. “On Optimal Decision-Making in Brains and Social Insect Colonies.” *The Royal Society. Journal of the Royal Society Interface* 6(40): 1065-74.

McVay, Jennifer C., and Michael J. Kane. 2009. “Conducting the Train of Thought: Working Memory Capacity, Goal Neglect, and Mind Wandering in an Executive-Control Task.” *Journal of Experimental Psychology. Learning, Memory, and Cognition* 35(1): 196-204.

Omohundro, Stephen M. 2008. “The Basic AI Drives.” In *Artificial General Inteligence* 171: 483-92.

Parthemore, Joel, and Blay Whitby. 2013. “What Makes Any Agent a Moral Agent? Reflections on Machine Consciousness and Moral Agency.” *International Journal of Machine Consciousness* 5(2): 105-29.

Pietrzykowski, T. 2018. “Personhood beyond Humanism: Animals, Chimeras, Autonomous Agents and the Law.” https://link.springer.com/content/pdf/10.1007/978-3-319-78881-4.pdf.

Schlosser, Markus. 2019. “Agency.” In *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/agency/.

Shing, Yee Lee, Garvin Brod, and Andrea Greve. 2023. “Prediction Error and Memory across the Lifespan.” *Neuroscience and Biobehavioral Reviews* 155 (105462): 105462.

Synofzik, Matthis, Gottfried Vosgerau, and Albert Newen. 2008. “Beyond the Comparator Model: A Multifactorial Two-Step Account of Agency.” *Consciousness and Cognition* 17(1): 219-39.

Tait, Izak. 2024. “Structures of the Sense of Self: Attributes and Qualities That Are Necessary for the ‘Self.’” *Symposion: Theoretical and Applied Inquiries in Philosophy and Social Sciences* 11(1): 77-98.

Taylor, Charles. 1985. “The Concept of a Person.” In *Philosophical Papers, Volume 1: Human Agency and Language* 97-114.

Turner, Alexander Matt, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2019. “Optimal Policies Tend to Seek Power.” *arXiv [cs.AI]*. arXiv.

Warren, William H. 1990. “The Perception-Action Coupling.” In *Sensory-Motor Organizations and Development in Infancy and Early Childhood*, edited by H. Bloch and Bertenthal, B.I., 23-37. Dordrecht: Springer Netherlands.

———. 2006. “The Dynamics of Perception and Action.” *Psychological Review* 113(2): 358-89.

Wynter, Adrian de. 2024. “Will GPT-4 Run DOOM?” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2403.05468.

Yaffe, Gideon. 2004. “Conditional Intent and Mens Rea.” *Legal Theory* 10(4): 273-310.