Intentional State-Ascription in Multi-Agent Systems

A Case Study in Unmanned Underwater Vehicles

Justin Horn, Nicodemus Hallin, Hossein Taheri,
Michael O'Rourke, and Dean Edwards

1 Introduction

Recently, considerable attention in AI research has been paid to multi-agent systems, or systems that comprise multiple intelligent or semi-intelligent agents interacting with one another. Agents in multi-agent systems are regularly described

Justin Horn
Department of Philosophy
Arts 2 Building,
18 Symonds Street,
Auckland
New Zealand

Nicodemus Hallin · Hossein Taheri
Mechanical Engineering Department
University of Idaho
Moscow, Idaho 83844-1021
USA

Michael O'Rourke
Department of Philosophy
University of Idaho
Moscow, ID 83844-3016
USA

Dean Edwards
Chemical Engineering Department
University of Idaho
Moscow, Idaho 83844-1021
USA

V.C. Müller (Ed.): Philosophy and Theory of Artificial Intelligence, SAPERE 5, pp. 225–235.
springerlink.com

using the language of *intentional states*, or states which refer to or are about something outside themselves. Examples of intentional states include, but are not limited to, goals, beliefs and desires.

How seriously are we to take these ascriptions of intentional states? Are members of multi-agent systems "true believers" in the sense that their intentionality is more robust, or are our ascriptions of intentionality merely a convenience of discourse that should not be given much weight? These questions frame the present agenda of the authors, who defend a version of the former position.

Our goal is to establish, through detailed examination of a case study, that multi-agent architectures embed the need to adopt the intentional stance toward them. This case study draws on work done by the University of Idaho's UUV (Unmanned Underwater Vehicle) research team, whose UUVs comprise a reasonably typical multi-agent system. The strategy is to develop conclusions which can be generalized to apply to many multi-agent systems, but which are also firmly rooted in the specific details of our case study. Bearing this in mind, the characteristics of the UUVs which ultimately lead the authors to support attribution of intentional states are characteristics the UUV fleet shares with many other multi-agent architectures. As we move forward, we will primarily focus on establishing our claims with respect to our case study, saving broader generalizations about other multi-agent systems for the final section.

2 Background

In "True Believers: The Intentional Stance and Why It Works", Daniel Dennett outlines a certain predictive strategy he calls "adopting the *intentional stance*" (Dennett 1997, 59). There are many sorts of stances we can adopt with respect to predicting the behavior of some object or system; adopting one of these stances amounts to highlighting one among a hierarchical stratification of conceptual levels at which processes take place. Dennett identifies the *physical stance*, at which we are concerned with the basic action of physical laws; this is the stance we might appropriately adopt with respect to the prediction of billiard balls. There is also the *design stance*, in which the object or system is conceived of as designed, i.e. having a purposive function. This would be a stance appropriate to adopt when predicting the behavior of, say, a wristwatch. We would expect, for example, that the second hand will complete one revolution around the face of the watch per minute, because its function is to allow its user to accurately gauge the passage of time.

Dennett then goes on to characterize the *intentional stance*, on which we interpret the object or system in question as an goal-directed agent:

> *Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the*

> *same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many - but not all - instances yield a decision about what the agent ought to do; that is what you predict the agent will do.* (Dennett 1997, 61)

Dennett's main points include the following. First, it is perfectly legitimate to ascribe intentional terms like belief, desire, goal, plan, and the like to objects and systems, insofar as adopting the intentional stance towards those objects and systems is appropriate, that is explanatorily or predictively fruitful. Second, it is impossible for one to avoid self-ascribing the intentional stance, and it is also impossible to avoid adopting it towards "one's fellows *if* one intends, for instance, to learn what they know." (Dennett 1997, 71).

With respect to the UUVs that compose the University of Idaho UUV fleet, we establish the following: (1) The UUVs, on the grounds of intercommunication, hypothetical reasoning, and mutual interest in each others' available information, can and in fact do adopt the intentional stance with regard to each other and themselves. (2) The behavior of UUVs is best understood (indeed, only fully understood) *by us* when we adopt the intentional stance toward UUVs. This is in part a consequence of the UUV design team manifestly adopting the intentional stance with respect to UUVs as a solution to hypothesized and encountered mission difficulties. If this argument is successful, and Dennett is right in maintaining that any intentional system will be an appropriate candidate for intentional state-ascription, then UUVs (and, consequently, other agents that belong to sufficiently similar multi-agent architectures) are appropriately seen as intentional agents.

However, as previously argued in Ray et al., we also have reason to conceive of the fleet as a whole as an intentional system, this would mean the fleet too would be considered an intentional agent, itself made up of intentional agents. Some might consider this a problematic or even self-refuting view. We argue to the contrary, pointing out three counter-objections. First, that we humans ourselves are composed of parts, at least some of which are most usefully predicted by adopting the intentional stance; we also compose larger social systems that are similarly best understood on the intentional stance. Second, that it is perfectly consistent to maintain that systems can have beliefs without their being *aware* of their having these beliefs; we regularly hold this view with respect to many types of lesser intelligent animals. Finally, the whole reason we are in the business of belief-ascription in the first place is so that we can accurately and economically predict behavior under different circumstances. These considerations all lead to the conclusion that UUVs and UUV fleets are to be included (albeit in their proportionally restricted degree), alongside ourselves and all other intentional systems, among the ranks of "true believers".

3 Inter-Agent Intentional State-Ascription

We shall begin by considering whether or not the University of Idaho's UUVs can reasonably be seen to interpret themselves and their fellow UUVs by adopting the intentional stance; we maintain that they can and do. To motivate this position, let us turn briefly to Hallin et al., in which the authors discuss the conditions that justify viewing some object or agent as appropriately "autonomous":

> *"An artificial system functions autonomously when its behavior is under its own control, or more precisely, when the system makes decisions concerning its own behavior that are not choreographed down to the last detail in advance and are responsive to changes in circumstance. To be responsive and in control, the system must allow new information as input to influence system output, where this influence is controlled by an information management infrastructure. In systems that communicate, such as the UI UUV fleet, this infrastructure will include a communication language and associated interpretation logics. The information management infrastructure is responsible for structuring the system's actual I/O (input/output) behavior, and...this infrastructure can be harnessed and put to use in planning for contingencies that could arise in the course of system operation."* (Hallin, et al. 2009, 2)

These UUVs work collaboratively to achieve a common mission goal, e.g. the detection of mine-like objects (MLOs) in a minefield, or analysis of a target ship's magnetic signature. In the course of these missions and simulations thereof, the UUVs engage in intercommunication and hypothetical reasoning, and they have a mutual interest in knowing what information is available to the other UUVs in the fleet. I argue that these considerations weigh in favor of the position that the UUVs regard one another as intentional agents, that is agents who have beliefs and goals, and who act on the basis of those beliefs to achieve those goals.

Let us begin with intercommunication. As noted in the quotation above, the UUV's send messages to one another, using AUVish, a language comprising 13-bit messages designed for the UUVs (Rajala, O'Rourke and Edwards 2006). In the context of a mine-countermeasure mission (MCM), the UUVs send messages containing information about which UUV is speaking, the role of that UUV in the fleet, and information about that UUV's current task assignment (e.g. "swimming in formation", "inspecting an MLO", etc.). Sometimes the messages go beyond mere reports; they can include, for example, a "request for permission to broadcast" a more detailed 32-byte message about, say, the location of an MLO. The implicit assumption here is that the UUVs expect the other UUVs to understand

the content of these messages as they do, and modulate their behavior appropriately on the basis of the messages intentional content.

AUVish messages contain intentional content; they are "about" the UUVs that send them, and in some cases they are "about" the shared environment in which the UUVs are operating. The UUVs select the messages that they choose to send on the basis of the interaction with their environment, and their behavior is modulated on the basis of which messages they receive. This intentional content and the way it modulates UUV behavior cannot be fully understood without reference to the representational content contained in these messages; this means that Dennett's condition, that there be predictive and explanatory usefulness of one UUV adopting the intentional stance toward another, is fulfilled.

An important parallel between the UUVs' intercommunication and the intercommunication of agents whose intentional status is less questionable (e.g. human beings) is that, like us, the UUVs can make mistakes. Because they are not infallible, it is necessary for the UUVs to distinguish between "the facts" (even if this is just a view of the facts from that UUV's perspective) and "the beliefs of the message sender". Also, sometimes messages get "lost in the shuffle", either due to technical failure or the intervention of environmental noise. In these situations, hypothetical reasoning is employed to correct error or maximize the fleet's efficiency in future actions. Using a Language Centered Intelligence (LCI) module, a UUV can generate hypotheses about future, present, or past scenarios by drawing conclusions based on the combination of information about the environment currently available to the UUV and other, hypothetical or counterfactual information about scenarios that may come to be or information that the UUV might be presently mistaken about (Hallin, et al. 2009). For example, a UUV might run through alternative power replacement scenarios if a battery is running low, or it might project anticipated messages from other UUVs for substitution in the event of an incomplete or missing message. The projection of hypothetical scenarios suggests that the UUVs must make a distinction between "the facts" and "beliefs" in their own case as well. Were there no such distinction, the UUVs would have no principled reason to act on some pieces of information but not on others. This underwrites self-ascription of the intentional stance on behalf of the UUVs.

Finally, and perhaps obviously, UUVs, have a mutual interest in the information available to the other members of the fleet. Information available to one UUV may not be immediately available to other members of the fleet. Collecting and synthesizing this body of information and tracking changes made to it in real time is crucial to the success of UUV missions. Also, as noted above, UUVs have a vested interest in tracking errors or discrepancies in this body of information, as these present obstacles to efficient and successful mission completion. As Dennett points out, if the UUVs want to "learn what their fellows know (or believe)", they must attribute the intentional stance to one another, and to themselves.

4 External Intentional State-Ascription

But what about how *we* regard agents in a multi-agent system? Might all this talk of UUVs intercommunicating about their knowledge and beliefs just be too fast

and loose? Do they *really* believe? Laying aside the question of what "real" belief consists in for now, let us consider an objection on which we try to avoid adopting the intentional stance toward the UUVs, adopting instead the physical or design stances. Adopting the physical stance here is borderline ridiculous. The kinds of interactions that are going on are too complicated and on much too large a scale to make the physical calculations practically tractable. Working out electron interchanges in one of the UUV circuit boards, for example, is just far too cumbersome to be undertaken, especially when more fruitful stances (i.e., design, intentional) are available. So what about the design stance? Well, part of the problem here is that, given the autonomous nature of UUVs as described above, the UUVs were *designed to be intentional systems!* From the very beginning, designers have approached the challenges presented by various missions with strategies that explicitly make use of the notions that UUVs are agents with beliefs and goals who interact with their environment and each other in light of these. Thus, an attempt on our part to adopt the design stance collapses into adopting the intentional stance. Given that the physical stance is a non-option, adopting the intentional stance with regard to UUVs is the only option we have left.

Perhaps we might argue that the artificial nature of the UUVs is grounds for withholding intentional status from them. Adams and Aizawa argue that "cognition involves particular kinds of processes involving non-derived representations" (Adams and Aizawa 2001, 53). Perhaps the fact that we bestowed the UUVs with the proper sort of structure to use the language they do, their representations are derivative, parasitic upon *our* non-derived representations, and thus UUVs are not properly possessed of mental states like beliefs. But we must be careful to avoid organocentrism here. To make this point clear, consider what we would say about a designed robot that had a silicon hardware unit that was a perfect functional model of an actual human brain. On what non-question-begging grounds could we deny that this robot properly held beliefs? So it cannot be a matter of medium or of having a designer that is the "mark of the cognitive".

In any case, it seems perfectly reasonable to see the UUVs' representations as arising within them, without our mediation beyond its design. This is, again, tied up with the conditions of their autonomy. The UUVs must mediate different sources of information in a complex environment (e.g. position, sensor information, incoming messages, mission time, etc.) with its own evaluative resources. It must compare and evaluate different possible courses of action with respect to multiple competing criteria, and then select from among these the option that will maximize the chance of efficient and successful mission completion. Given that they do all this *on their own* in the field and in simulation, it seems appropriate to identify the UUVs as the source of their own representations, undermining the objection at hand.

5 Relationship between Collective and Individual Intentionality

While we are considering objections to this position, let us spend some time on a very different type of objection to this view. This objection turns on the idea that

a part of an intentional system cannot itself be an intentional system.[1] This is a view we will ultimately reject, but before doing so, we should outline the view as it might be defended.

Consider yourself. You are, undoubtedly, an intentional system. You have beliefs, desires, goals, and many other types of mental states infused with intentional content. Say you believe that Stevie Wonder is a great musician; no problems so far. Now consider some part of you, say, your left hand. Can your left hand believe that Stevie Wonder is a great musician? No, that doesn't seem right. But maybe we are looking at the wrong type of part here; what about your brain? Does it make sense to say that your brain believes that Stevie Wonder is a great musician? This also seems like a potential category mistake. Brains don't *have* beliefs, they are where the brain-haver's beliefs are stored, or physically located, or some such thing. Compare "I am thinking" with "My brain is thinking"—this phrasing seems awkward or uncomfortable at best. I suggest that this awkwardness is what motivates the objection we are about to consider.

Ray et al. argue in "The Ontological Status of Autonomous Underwater Vehicle Fleets" that we ought to accord agent-status to *the fleet* of UUVs. Because of the complexity of the missions undertaken by UUV fleets, there are some complex patterns of actions that cannot be made sense of without the postulation of the fleet as a single entity; that is to say, *emergent behavior* arises, behavior that cannot be reduced to the aggregate sum of collective behaviors. Ray et al.'s discussion of ant colonies is illustrative:

> *"...multiple agents acting collectively are capable of performing certain actions that cannot be reduced to the actions of multiple agents acting individually. Examples of this type of emergent behavior include ant colony relocation and evasive herd movement. Ant colonies are generally thought to behave as a single entity rather than as a mere aggregate of individuals. This is due to the fact that there are certain things an ant colony, and only an ant colony, can do, e.g., relocate and nurture the queen ant. In fact, there is an entire class of predicates reserved for the ant colony itself."* (Ray et al., 2009)

The idea here is that if we see the UUV fleet as more ontologically important than the individual UUVs, then the UUVs considered individually will just be a part of the fleet. And if it is the case that the *fleet* is intentional, and the individual UUVs are just parts of that, it will be hard to see them as candidates for proper belief-ascription for the same reasons we are intuitively uneasy about ascribing intentional status to mere parts of ourselves.

If we accept that we should accord ontological priority to the fleet, what reasons do we have for seeing that fleet as itself an intentional agent? This very

[1] Excepting, of course, the part that is identical with the system as a whole.

question frames the discussion in Ray et al.'s "Using Collective Intentionality to Model Fleets of Autonomous Underwater Vehicles". There, the claim that fleets should be collectively afforded intentional status is extensively defended. Ray et al. characterize collective intentionality in the following way:

> *"Collective intentionality is exhibited by a group of agents that pursues a goal as a group, exploiting distributed states that are jointly directed at the goal. This type of intentionality involves goal directed behavior that is irreducibly performed by the fleet itself and so is not simply the sum of individual vehicle actions. Searching a given space and generating a map would be an example of an irreducibly goal directed behavior...since it involves distributed processing and information gathering. The generation of a map is only possible insofar as the vehicles cooperate with each other and exchange information necessary for the generation of a map."* (Ray et al., 2009)

So, we have our objection by double syllogism. Parts of properly intentional agents or systems aren't themselves properly intentional, a UUV is a part of a UUV fleet, and UUV fleets are properly intentional agents or systems. Therefore, parts of UUV fleets aren't themselves properly intentional, and as this applies to UUVs (being parts of UUV fleets), UUVs are therefore not candidates for proper belief-ascription. We try to meet this objection, by rejecting the initial supposition that parts of intentional systems or agents cannot themselves be intentional.

We might begin by pointing out that the fact that just because many parts of us don't constitute properly intentional systems doesn't mean it couldn't happen in other cases of intentionality. No *necessary* connection has been established; this might be an accidental feature of the way intentionality is realized in us. However, we would like to go farther and suggest that there are at least some parts of human beings properly understood on the intentional stance. Consider the human immune system. The immune system traffics in information that is not readily available to us as human agents in the same way that our perceptual information, for example, is readily available. The immune system can be seen as representing information about objects it encounters in the body, and can be seen as taking specific action on the basis of that information. Furthermore, this activity is goal directed, attempting to restore your body to an "equilibrium" of health. Now it certainly seems right to say that one's immune system can do things that are beyond one's control or often even one's awareness, say, increasing blood flow to a particular area in the body. It seems to make more sense to ascribe these actions to the immune system than it does to ascribe them to me as a conscious agent. "I didn't increase the blood flow to my leg; my immune system did that!" But clearly, my immune system is a proper part of me. So here we have a counterexample to the thesis that a part of an intentional system cannot itself be intentional.

Also, we as humans make up multi-agent systems that are themselves collectively intentional. If the notion of collective intentionality makes sense with respect to an artificially constructed UUV fleet, then surely it must apply to groups of humans, possessed of their own individual intentionality. Consideration of such groups of humans is (among other things) what gave rise to the idea of collective intentionality in the first place! Football teams huddle around a quarterback or try to counter a blitz. Nations war with and invade other nations. Orchestras play symphonic works or accompany soloists. If you are inclined to accept the idea of collective intentionality (which is required for the objection to go through), then certainly all these types of groups exhibit it as well, and they do so without threatening the individual intentional capacities of the constituent members. To the contrary, it would seem the collective intentionality supervenes on the intentionality of the members, the state of the collective being determined by but not identical with the intentional states of the members.

We should remind ourselves here that there is no contradiction in maintaining that systems can have beliefs without their being *aware* of their having these beliefs; we regularly hold this view with respect to many types of lesser intelligent animals. Self-consciousness is not a prerequisite for belief, or intentional status in general. Dogs know where they buried a bone in the backyard. Bees transmit information to their fellows about the location of pollen sources. Dogs and bees, then, have beliefs, or at least intentional states, but it is not clear that dogs are *aware that they have beliefs*; it is almost certain that bees are so unaware. Again, we must avoid the pitfall of over-generalizing accidental features of our own cognitive profile.

Finally, we should look at the role belief-ascription plays for us. What good does it do for us to ascribe beliefs to others? Why aren't we all solipsists, especially given our lack of ability to access the beliefs of others in the way we access our own? The whole reason we are in the business of belief-ascription in the first place is so that we can accurately and economically predict behavior under different circumstances. If I attribute beliefs to you, it helps me to understand your behavior in ways that are not available without the resources of intentionality. This point is echoed in McCarthy's discussion of appropriate conditions for intentional-state ascription:

> *"To ascribe beliefs, free will, intentions, consciousness, abilities, or wants to a machine is legitimate when such an ascription expresses the same information about the machine that it expresses about a person. It is useful when the ascription helps us understand the*
>
> *structure of the machine, its past or future behavior, or how to repair or improve it. It is perhaps never logically required even for humans, but expressing reasonably briefly what is actually known about the state of the machine in a particular situation may require mental qualities or qualities isomorphic to them."* (McCarthy, 1979)

In this quotation, McCarthy points out that we are not even *forced* to ascribe intentional states to other humans. We generally do so because of what these ascriptions *buy* us. If this justification is sufficient to underwrite appropriate intentional state-ascriptions to other people, then it should be sufficient in cases of non-human multi-agent systems as well. Given this, even if your intuitions still pull you strongly in rejecting the intentionality of anything non-human, you should consider the ways in which your belief-ascription helps you predict behavior in this domain, and the ways in which it *could* help you predict the behavior of other agents and systems, should you be able to overcome your anthropocentrism.

6 Conclusions

We are now in a better position to see how our conclusions with respect to the UUVs and the UUV fleet can be generalized to other multi-agent systems. While the University of Idaho's UUV fleet is concerned with performing very specialized "niche" tasks, almost none of the specific details of these tasks are necessary to establish our conclusions. Rather, our conclusions are based on two general features of the fleet architecture that it shares in common with many other multi-agent systems. First, that the members of the system, in the course of typical actions in their environment, must engage in processes which attribute intentional states to themselves and/or one another in order to "get the job done". In the UUV fleet, these processes include vehicular intercommunication and hypothetical reasoning, but other sorts of processes might fit the bill as well, so long as they traffic in intentional states. The second feature (which perhaps dovetails with the first) is that the system *was designed to be an intentional system*. It is this fact which, in our case study, removes the possibility of rejecting the intentional stance in favor of the design stance, as the latter collapses into the former. Thus, we expect that anyone who is convinced by the arguments we have presented with respect to our case study will be similarly inclined to accept parallel conclusions about other multi-agent systems that exhibit these two features.

So if we are to accept that all sorts of individuals and groups of them *are* intentional, are they all intentional in just the way that we are? To the degree that we are? In conclusion, we offer a viewpoint which, while according some non-human agents and systems "true believer" status, this is mitigated by a reduced richness of belief as complexity of the system decreases. This is a sort of "sliding scale" approach, on which intentionality and beliefs are "thick" concepts. That is, one can be intentional, or have beliefs, to a greater or lesser degree; there are many figurative "shades of grey" between the black-and-white extremes of full-on belief having (like ours) and total lack of belief (like a stone). The complexity of the system, in its sensitivity to different types of information, its ability to represent non-actual states of affairs, and the range of actions with which it can respond, will be correlated with the richness of intentionality, or the seriousness with which we take the ascription of belief.

In support of this idea, let us look back one more time at our near and more distant relatives across the animal kingdom. We might organize them into a kind of "cognitive hierarchy", with microbes and sea slugs near the bottom, insects a little

further up, lizards, birds, and eventually mammals, topping out with perhaps dolphins and chimpanzees (and maybe an octopus) and finally humans. The details of who fits in exactly what slot may be contentious, but the idea that slugs aren't as smart as dogs, who aren't as smart as us, shouldn't be. But now we have the beginnings of a sort of cognitive sorites series: a gradual increasing or decreasing of cognitive status on a sliding scale. Now, we may be tempted to try and draw a cognitive "line in the sand" somewhere, between the believers and the non-believers. The problem with this (as with other sorites series) is that there is no non-arbitrary way to decide where to draw such a line. The best solution is to reject the idea that belief is an all or nothing affair; rather, it is a matter of degree.

So, in light of this, the recommended position is to see both UUVs and UUV fleets (and, correspondingly, many multi-agent systems) as legitimately intentional or collectively intentional agents or systems, respectively. However, given our increased complexity and nuance of informational and behavioral modulation, we humans believe "more richly" than any artificial agents are currently able to. This is a win-win; humans retain an elevated status as the richest and most intentional believers (at least for the time being), and UUVs, UUV fleets, and other non-human agents in multi-agent systems are accorded status as real, legitimate believers, albeit in their proportionally reduced degree.

References

Adams, F., Aizawa, K.: The bounds of cognition. Philosophical Psychology, 43–64 (2001)

Dennett, D.: True Believers: The Intentional Stance and Why It Works. In: Hagueland, J. (ed.) Mind Design II: Philosophy, Psychology, and Artificial Intelligence, pp. 57–79. MIT Press (1997)

Hallin, N., Egbo, H., Ray, P., O'Rourke, M., Edwards, D.: Enabling Unmanned Underwater Vehicles to Reason Hypothetically. In: Proceedings of Oceans 2009 MTS/IEEE Biloxi, Biloxi, Mississippi (2009)

McCarthy, J.: Ascribing Mental Qualities to Machines. Technical Report, Stanford AI Lab (1978)

Rajala, A., O'Rourke, M., Edwards, D.: AUVish: An Application-Based Language for Cooperating AUVs. Oceans 2006 (2006)

Ray, P., O'Rourke, M., Edwards, D.: The Ontological Status of Autonomous Underwater Vehicle Fleets. In: Proceedings of Oceans 2009 MTS/IEEE Biloxi, Biloxi, Mississippi (2009)

Ray, P., O'Rourke, M., Edwards, D.: Using Collective Intentionality to Model Fleets of Autonomous Underwater Vehicles. In: Proceedings of Oceans 2009 MTS/IEEE Biloxi, Biloxi, Mississippi (2009)