Loop-closing semantics

Ian Wright

Abstract How is semantic content possible? How can parts of the world refer to other parts? On what grounds (if any) can we claim that simple mechanisms, such as thermometers, thermostats, clocks and rulers etc., refer to features of the world in virtue of their causal powers rather than our intentional practices with respect to them? I introduce Sloman's Tarskian-inspired 'loop-closing theory' in order to answer these questions. Loop-closing theory reduces a subset of semantic properties to the causal properties of control systems. I develop Sloman's acccount by specifying a metalanguage to describe the causal structure of loop-closing models, and then identify and define a control system's *manipulable feature*, which is the subset of the world necessarily present for control success. Loop-closing theory identifies the referential content of a control system's information-bearing substates with the manipulable feature. I conclude by applying loop-closing semantics to some illustrative test cases, such as the semantic properties of memory addressing in CPUs, the referential content of bacterial magnetosomes, the problem of misrepresentation, and connections to Ramsay-Whyte success semantics.

> 'In principle, if you want to explain or understand anything in human behavior, you are always dealing with total circuits, completed circuits. This is the elementary cybernetic thought.'
>
> Gregory Bateson (1999)

1 Semantic properties

I believe the temperature of the room is 20 °C. Clearly my belief is not the same thing as the temperature of the room. My belief *refers* to something else – it has semantic content.

Ian Wright
Open University, Economics, Faculty of Social Sciences, Milton Keynes, UK, e-mail: wrighti@acm.org

My belief picks out some isolated feature or features of the world, while excluding all others. For example, my belief does not refer to air pressure. So the semantic content is *focused* and often univocal.

And I may be mistaken. My beliefs have truth conditions; they are defeasible. For example, it just so happens that the room temperature is 15 °C. My belief has a *semantic value*, which here is 'false'.

Can these semantic properties – reference, focus and semantic value – be reductively explained in terms of a theory that does not presuppose semantic properties? Can intentional phenomena be reduced to known non-intentional phenomena and therefore naturalized?

In this paper I address this question with reference to simple artificial artifacts, especially thermometers and thermostats. These simple devices are *prima facie* unpromising candidates for admission into the class of things that (non-derivatively) possess intentional properties. But unlike minds they possess no 'black box' mysteries that force us to conjecture about how they really work. Since they are so simple they can be understood in every detail. Hence, if we demonstrate that some simple artifacts in fact possess intentional properties, independent of our practices, then we will have a clear understanding of exactly how (some) intentional properties are reducible to non-intentional properties.

Although I make extensive use of simple examples the ultimate explanatory target is, of course, cognition in general – whether simple, complex, natural or artificial.

First, I consider a 'straw man' theory of semantic properties, and point out its problems, in order to introduce the explanatory requirements that any account of semantic properties must satisfy.

2 A crude causal theory

Consider a thermometer that indicates temperature by the height of its mercury column, which is suspended in a capillary tube. As the local temperature rises, the mercury expands; when the temperature falls, it contracts. The column height is calibrated against a standard scale (e.g., °C), which we read off.

A necessary condition for a thermometer to measure temperature is the lawful covariation of some part of it with temperature. For example, the design of the thermometer exploits the natural law of thermal expansion. The height of the mercury column is an 'information-bearing sub-state' (Sloman, 1994b), or 'representational vehicle', that reliably indicates the local temperature. We might propose, then, that the substate represents temperature in virtue of this lawful covariation.

This proposal forms the theoretical core of a family of 'information semantic' (Fodor, 1990, Ch. 3) theories of content. Typical examples are Dretske (1981), Fodor (1990), Jacob (1997) and Barwise and Seligman (1997) (although, I should stress, each of these examples considerably extends and alters this core and is not reducible

to it). Information semantic theories maintain, in one form or another, this initial premise:

(IS) Information-bearing sub-state X refers to Y if Y reliably causes the state of X.

For example, the height of the mercury column refers to temperature because temperature reliably causes its height. The precise specification of 'reliable cause' differs between theories; it may be expressed in terms of conditional probabilities, 'nomic regularities' (laws), or sets of counterfactual dispositions etc.

Some mental representations are not obviously caused by what they refer to (e.g., imagining flying pigs) but (IS) makes no claim to be a complete theory of representation.

Let's now examine the problems with this 'crude causal theory' (Fodor, 1989, ch. 4).

3 Conjunction problems

Consider this complicated setup. Submerge a heating element in a container filled with a liquid. Float a small thermo-kinetic engine on the liquid's surface. Connect the engine to a pulley system that opens or closes an aperture connected to an independent source of hot air. Place a thermometer before the aperture. Now turn on the heating element. As the liquid's temperature rises the heat induces rotary motion in the engine, which opens the aperture blowing hot air onto the thermometer. (In general, we can imagine arbitrary 'Rube Goldberg' machines in-between the object whose temperature is to be measured and the thermometer).

In this specific setup the state of the thermometer's information-bearing sub-state is (at least) reliably caused by the output of the heating element, the liquid's temperature, the speed of the engine, the size of the aperture and the amount of hot air flowing over the thermometer. According to (IS) the height of the mercury column represents the output of the heating element *and* the liquid's temperature *and* an engine speed *and* the size of an aperture *and* the flow of hot air, or any subset of this conjunction.

The point is this: a reliable causal chain may be complex and therefore support multiple 'upstream' candidates for the semantic content of any 'downstream' information-bearing sub-state in that chain. (IS) picks out the *conjunction* of all the antecedent covarying features as the semantic content of the thermometer. The thermometer's information-bearing sub-state is therefore semantically indeterminate: we have a 'conjunction problem'. So (IS) has difficulty explaining semantic focus since it provides no basis – independent of our intentional practices – for claiming that a thermometer represents (only) temperature.

Rube Goldberg machines are highly contrived. Perhaps we should exclude such unusual situations. In normal circumstances, most of the time, temperature is the only feature of the world that reliably causes the height of the mercury column. Can we therefore reject the conjunction problem on these grounds?

No. The premise is false. Temperature is not the only feature that reliably causes the height of a mercury column. For example, a thermometer that measures air temperature also normally measures air pressure (Jacob, 1997, pg. 103) since temperature and pressure are properties that often covary (Gay-Lussac's law). And a thermometer that measures the temperature of a liquid also normally measures the rate of evaporation of molecules from the surface of the liquid.

Rube Goldberg machines, therefore, are not required to expose the conjunction problem. In fact, highly contrived, and therefore unusual, experimental situations – such as a thermometer measuring the temperature of the air inside a balloon or the temperature of a completely sealed liquid – are necessary to exclude air pressure and evaporation as covarying properties. We normally ignore these covarying features when thinking of the function of a thermometer but a reductive theory of semantic properties cannot.

Perhaps we can exclude distal features of the causal chain – color, engine speed, aperture size etc. – by restricting (IS) to the *proximal* feature, or final cause, of the height of the mercury column? The final cause, in our example, is hot air blown across the thermometer. Such a modified theory would pick out temperature (and not color, engine speed or aperture size).

But this restriction fails to generalize since the final cause of a thermometer's reading need not be an instance of the natural kind 'temperature'. For example, infrared thermometers, widely available, measure temperature at a distance by focusing infrared light from the measured object to an internal detector. The proximal feature, in this case, is not temperature but light. (Furthermore, it's nonsensical to to claim that when I read a nearby thermometer, and form the belief that the room temperature is 20 °C, then my belief, in fact, refers to the proximal projection of light on the surface of my retina.) The distinction between distal and proximal features doesn't solve the conjunction problem.

4 Disjunction problems

Beliefs can be in error. They have semantic values, such as 'true' of 'false'. So any reductive theory of semantic content needs to explain how a representation can misrepresent, or fail to refer.

Place a mercury thermometer in a bowl of dry ice that is below -40 °C. The mercury freezes solid. In consequence, the thermometer registers the wrong temperature, say -35 °C, and the semantic value of its information-bearing sub-state is false.

(IS) claims that the height of the mercury column represents temperature because temperature reliably causes it. At very low temperatures mercury freezes solid and disrupts this causal relationship. At first glance, therefore, (IS) appears to successfully explain misrepresentation.

But (IS) appears to work only because we (almost reflexively) foreground circumstances in which the thermometer functions as intended from circumstances in

which it does not. But our beliefs about 'normal circumstances' don't figure in (IS). In fact, the height of *solid* mercury is also reliably caused by temperature with the difference that the coefficient of thermal expansion is about half that of its liquid state. Hence, according to (IS), the height of the solid mercury column *also* represents temperature and therefore this is not a case of misrepresentation. It just so happens that the manufacturer's calibration of the thermometer assumes that normal circumstances prevail.

Let's consider another case. Place a mercury thermometer in a microwave oven and switch it on. The microwave radiation induces an electric current in the mercury column, which heats it. The mercury column expands until it vaporizes at 350 °C. The microwaves don't directly heat the air inside the oven. Hence, in these circumstances, the thermometer fails to measure ambient temperature. Is this a case of misrepresentation?

The thermometer is not being used as intended by its designers. But (IS) is independent of the intentions of designers. In this setup the height of the mercury column is reliably caused by the temperature of the mercury or, conjunctively, by the duration the oven has been switched on, or with the integral of the microwave energy supplied etc. Hence, according to (IS), the thermometer's mercury column may correctly represent any of these features.

Fodor (1990, pg. 42) identifies the general problem: if representation X refers to Y in virtue of Y reliably causing X then it follows that X 'reliably' or 'truthfully' represents Y. But a theory of semantic content should be able to explain the possibility of representational error and, therefore, the conditions for X to be a representation of Y must be somehow *separable* from the conditions for X to be a veridical representation of Y. In (IS) these conditions are identical and therefore (IS) cannot account for representational error.

Purported examples of misrepresentation turn out to be cases of representation according to (IS). So all the different situations in which a Y reliably causes X are a referent of X. (IS) says that X refers to Y_1 *or* Y_2 *or* Y_3 *or* Y_4 etc. – that is the *disjunction* of all the possible reliable causes of X in different circumstances.

The conjunction and disjunction problems are well-known classic counterarguments to (IS) that generate semantic indeterminacy. The conjunction problem highlights synchronic indeterminacy (i.e., indeterminacy in a given situation) and the disjunction problem highlights diachronic indeterminacy (i.e., indeterminacy due to multiple situations). In consequence (IS) admits too many candidates for semantic content and therefore fails to explain how reference is focused. The disjunction problem also entails that (IS) cannot explain misrepresentation.

Many candidate Ys are admitted by conjunction and disjunction problems. I now give a further, and novel, argument that demonstrates that (IS) also fails to provide sufficient reasons why a substate X refers to *any* of these Ys. In other words, (IS) does not even explain how referential content is possible.

5 The problem of semantic dualism

Consider again the mercury thermometer. First, let's check it's working. Stick it in a bowl of ice and wait for thermal equilibrium. The thermometer reads 0 °C. Now submerge it in a bowl of boiling water. The mercury column rises until it reaches 100 °C. All is well.

Now imagine a man from the moon who descends to Earth. He knows a little (moon) science but is entirely ignorant of human practices, language and notation. He finds an artifact lying on the ground (it's the thermometer we checked but he doesn't know that). The symbol '°C' written on its face means nothing (it looks like '*♣' to his eyes). He's curious and therefore starts to experiment on the artifact and, after hours of testing and study, he decides the artifact has something to do with temperature. So he sticks it in a bowl of ice. It reads 0 °C. He then submerges it in a bowl of boiling water. It reads 100 °C.

He now has the Aha! moment: he reasons that the symbol '°C' is a unit of *length* since it appears beside height marks. He conjectures that the Earth artifact is a device for measuring the coefficient of thermal expansion of mercury, which he calculates is $\frac{100}{x}$ °C/°M (that is, Earth units of length per moon units of temperature), where x °M is the difference between the boiling and freezing points of water measured in moon units of temperature. (And from an anthropological point-of-view he is delighted because now, via the objective properties of mercury, he can translate between Earth units of length, °C, and moon units of length!)

Where did the man from the moon go wrong? He didn't. His conjecture was consistent with the evidence. His only 'mistake' was to interpret the thermometer's scale to be a local measure of mercury height rather than an ambient measure of temperature. He didn't know that humans interpret the scale as *referring* to temperature.

The °C scale represents temperature when the artifact is used to measure temperature. But the °C scale represents length when the artifact is used to measure the thermal expansion of mercury. It just so happens that (unbeknown to the man from the moon) humans normally use the artifact as a thermometer and, in human notation, °C in fact means units of temperature.

Hence the relation of reliable causation between the local temperature and the height of the mercury column is *essentially* semantically indeterminate: the very same causal relation, in itself, is consistent with the mercury's height referring to temperature *and* referring to nothing, i.e. merely 'representing' itself. The fact that '°C' is conventionally interpreted as a unit of temperature depends on the intentions and practice of the thermometer's users. So the semantic properties of the thermometer are not reducible to reliable causation.

But perhaps this is too hasty? The man from the moon could have considered other factors, such as the thermometer's overall design. The bulb at the end of the capillary tube gives a clue to its true function. The mercury is sealed, which means the device could not be used to measure the thermal expansion of other liquids, which seems an odd limitation. Also, he discovered the artifact in a greenhouse filled with plants.

It's true: all this context, like pieces of a puzzle, helps identify the normal function of the artifact. Yet it would remain a fact that the artifact may be used to measure the thermal expansion of mercury. And when used in this way the height of the mercury column represents itself – and not something else.

In summary, the thermometer's information-bearing substate of a thermometer has at least two related, but distinct, semantic contents: indirect reference to temperature or direct reference to the height of the mercury column. It just so happens that our standard practice with respect to thermometers hides the direct reference. The identical relation of reliable causation between the thermometer and its circumstances is consistent with either semantic content. The conclusion follows: the semantic content of the thermometer's information-bearing sub-state, what 'it measures', is not determined by (IS). Hence there must be some other cause of semantic content not reducible to reliable causation.

Are mercury thermometers special in this respect? Consider a digital thermometer with a built-in platinum thermistor. The thermistor's electrical resistance covaries with temperature. An internal CPU samples the resistance and converts the data to °C for display on a LED screen. Again, assuming ignorance of human ways, does the number on the LED screen represent temperature or electrical resistance? The answer depends on whether the device gets used as a thermometer or as a means for determining platinum's 'temperature coefficient of resistance'.

Perhaps we can reverse-engineer the software code that converts resistance to '°C' in order to fix the semantics without reference to how the artifact is used? No because the code only reveals that electrical resistance is mapped to an output scale. But what that output scale represents is precisely the question. And if the manufacturer accidentally littered an obscure section of chip memory with some debug symbols, such as

```
float convertToTemperature( float resistance )
```

that would help decide normal use. But it would not alter the fact that the artifact can be used to measure platinum's temperature coefficient of resistance.

What's happening here? When we read a thermometer we 'look through' the height of the mercury column as if it were a transparent window onto temperature. But we can also 'look at' the mercury column as if it were an opaque state caused by temperature. The transparent and the opaque semantic contents are necessarily dual to each other. Let's call this 'semantic duality':

> (SD) Y is a reliable cause of the state of X. Then (IS) claims that X represents Y (transparent content). But X may also 'represent itself' as reliably caused by Y (opaque content) in virtue of the identical causal relation.

How general is (SD)? Consider a clock. Obviously the transparent content is time. The opaque content, for an analog clock, is the angle of rotation of the clock's hands. A clock is internally driven by its timekeeping element (e.g., pendulum, quartz crystal etc.) The clock's scale therefore represents time, if we use it to measure time, or angle of rotation, if we use it to measure the degree of hand rotation per tick (e.g., per swing, per oscillation etc.) of the timekeeping element.

Thermometers and clocks aren't that simple. Perhaps very simple measuring instruments lack a dual reading? Consider a ruler. The transparent content is the length of any adjacent object (a ceramic tile, say). The opaque content is the quantity of segments in the ruler's body. So whether the ruler's scale represents tile length or ruler segments depends on whether we want to measure the length of the tile or the quantity of ruler segments per tile.

Is this last example forced? No. Measuring the 'quantity of ruler segments per tile' is the kind of measurement that metrologists perform to calibrate rulers to a standard scale. For example, until recently, the 'meter' was defined as the length of a standard metal bar stored at constant temperature. So to calibrate and mark a two meter ruler a metrologist measures the 'quantity of ruler segments per metal bar', which, in this case, would be two segments. (Often the opaque semantic content of a measuring device is related to a calibration use-case.)

The examples can be multiplied endlessly (e.g., barometers, accelerometers, compasses, voltmeters, spirit levels, weather vanes, sundials etc.) (IS) cannot pick out what these devices actually measure because (SD) implies that a given relation of reliable causation supports both transparent and opaque content. In fact, whether a thermometer measures temperature (transparent content) or the height of a mercury column (opaque content) depends on the user's intentional choice to either measure temperature or the thermal expansion of mercury. The 'crude causal theory', therefore, does not break out of the 'intentional circle' and fails to explain why a substate X should possess transparent (i.e. truly referential) content at all.

6 Loop-closing semantics

Now that we understand some of the problems that any reductive account of semantic properties must face let's turn our attention to Sloman's loop-closing theory.

Tarski (1956) showed that an axiom system S, expressed in a formal language L, admits certain semantic interpretations while excluding others. Tarski's idea was to setup a mapping between formulae in L and structures in a domain of interpretation M. We may then interpret a formula F as 'referring' to or denoting a structure in M.

In Tarskian semantics F has the semantic value 'true' if it happens to denote a truth in M. The domain of interpretation, M, is therefore called a 'model' of an axiom system if each of the axioms denotes a truth in M. For example, the set of natural numbers is a 'model' for Peano's axioms of arithmetic.

In general, axioms system admit multiple different interpretations. As Hilbert famously remarked 'instead of points, lines, and planes one could consider an interpretation of geometry in terms of "tables, chairs and beer mugs"' (Feferman and Feferman, 2004, pg. 279). Hence, what a formula F possibly 'represents' depends on our intentional practice of setting up mappings between L and M. Tarski's theory is therefore not a reductive explanation of semantic properties (and was not intended to be). Nonetheless Sloman takes it as a starting point for constructing a reductive explanation of semantic content.

Sloman proposes we generalize Tarski's theory to (i) encompass non-formal 'languages' or representational systems, S, implemented as part of the cognitive machinery of an autonomous agent acting in a world; and (ii) consider the causal links between an agent (that uses S to control its actions) and its world. The main idea is that an agent's information-bearing substates S admit 'Tarskian' interpretations in terms of structures of an environment E (where S and E are analogous to a logical formula F and a structure in M in Tarskian semantics). But 'the existence of causal links removes, or reduces, the ambiguity inherent in the purely [Tarskian] structural semantics ... For example, electronic mechanisms ensure that bit-patterns in a computer are causally related to locations in its own memory rather than locations in another machine, despite having the same structural relations to both.' (Sloman, 1997).

Clearly, generalizing a highly-developed meta-mathematical theory to include all kinds of representational systems in dynamic interaction with all kinds of environments is no small task. In fact, it defines a whole research program. Sloman (1997) therefore offers us a 'thumbnail sketch', which is worth quoting in full since it's the essential programmatic statement of loop-closing semantics:

> Consider an environment E containing an agent A, whose functional architecture supports belief-like and desire-like sub-states. Suppose A uses similar sub-structures for both, just as a machine can use bit-patterns both for addresses and for instructions. Then we can define the class of possible 'loop-closing' models for a set of structures S by considering a set of possible environments E satisfying certain conditions, when the action-producing mechanisms, the sensors, and the correspondence tests are working normally:
>
> (a) States in E will tend to select certain instances of S for A's belief-like sub-states.
> (b) If Si is part of a desire-like state of A and E is in state Ei, A's correspondence tests show a discrepancy between Ei and Si, then (unless A's other belief-like and desire-like states interfere) A will tend to produce some environmental state Ej in E which tends to pass A's 'correspondence' test for Si.
> (c) If that happens Ej will tend to produce a new belief-like state in A.
>
> I repeatedly say 'tend to' to indicate that there are many additional factors that can interfere with the tendency, such as conflicts of desires, perceptual defects, accidents, wishful thinking, bad planning, and other common human failings. So these are very loose regularities, and cannot be taken to define internal states in any precise way. None of this presupposes that A is rational. It merely constitutes a partial specification of what belief-like and desire-like mean. However, a full specification will be relative to an architecture, within which functional roles can be defined more precisely.

A number of key ideas are worth noting in this passage.

First, a loop-closing model is an environment in which agent A 'tends to' achieve a goal. Internal structures S play two kinds of functional role during this episode. Belief-like states are caused by features in the world. Internal 'correspondence tests' compare desire-like states with belief-like states. Discrepancies cause the agent to act on the world until the tests are passed. A mind, on this account, is therefore a generalized feedback control system, a point Sloman repeatedly emphasizes (e.g., (Sloman, 2002b)).

Second, in Tarskian semantics the causal links between formula F and a structure in model M are instantiated by our (normally meta-mathematical) practices; but in

loop-closing semantics the links between sub-states S and environment E are instantiated by the agent's practical activity. The agent – as an autonomous, goal-directed mechanism – creates 'causal links' between its representations and referents. An independent observer, or 'truth-maker', such as ourselves, which sets up and maintains the mapping, is not required.

This thumbnail sketch, of necessity, leaves many questions unanswered. In particular, the sketch doesn't pick out which *subset* of 'states in E' are the semantic content of the agent's internal states. If the 'states in E' are a conjunction of features then loop-closing semantics is subject to the conjunction problem. Sloman observes that the causal links are 'loose regularities', perhaps due to 'perceptual defects', and hence the internal states cannot be defined 'in any precise way'. This suggests that loop-closing semantics is subject to the disjunction problem, and will therefore fail to identify cases of misrepresentation. We therefore need to add more detail to Sloman's sketch in order to understand how it avoids the kind of semantic indeterminacy that de-railed the 'crude causal theory' (IS).

As we develop Sloman's sketch I pay repeated attention to a very simple example of a feedback control system, a thermostat coupled to a heating system. This will help make some abstract ideas more concrete.

7 Thermostats

The thermostat is a familiar household device that embodies the principles of negative feedback control (e.g., Wiener ([1948] 1975, pg. 96–97)). Users set the thermostat's 'set-point' to the desired temperature. A thermometer-like component of the thermostat measures room temperature. The thermostat compares its set-point to the measured temperature. If the set-point is higher the thermostat turns heating on; but if the set-point is lower the thermostat turns heating off (or – if it has the capability – turns cooling on). The thermostat heats or cools the room until the temperature equals the set-point.

To be more concrete, we'll keep in mind a specific thermostat design where the thermometer component is a spiral bimetal coil, composed of two metals with differing coefficients of thermal expansion, which winds smaller or unwinds larger as the temperature changes. Moving the set point rotates the spur of the bimetal coil, which physically tips a bulb containing mercury. Mercury flows to one end of the bulb and completes the heating circuit. As room temperature rises the bimetal coil unwinds. Heating stops when the unwound coil untips the bulb and breaks the circuit.

McCarthy (1979) recounts the story of a heating engineer he called to his home.

> Recently it was too hot upstairs, and the question arose as to whether the upstairs thermostat mistakenly believed it was too cold upstairs or whether the furnace thermostat mistakenly believed the water was too cold. It turned out that neither mistake was made; the downstairs controller tried to turn on the flow of water but couldn't, because the valve was stuck. (McCarthy, 1979)

These intentional ascriptions – 'believed', 'tried', 'mistake' – are useful especially 'if we consider the class of possible thermostats, then the ascribed belief structure has greater constancy than the mechanisms for actually measuring and representing the temperature' (McCarthy, 1979). Dennett (1997) also emphasizes the utility of viewing the thermostat 'as if' it had belief-like and desire-like states that refer to temperature.

In what follows we seek to understand, not the utility, but the actuality of these ascriptions; that is, can loop-closing semantics answer the question of whether a thermostat's information-bearing substates *in fact* refer to temperature, regardless of what we may think, or how we talk about or interact with it. In other words, in contrast to Dennett's approach, we seek to identify how semantic properties necessarily emerge from certain classes of causal structures.

First, a preliminary assumption. I assume that the identification of a thermostat (and, more generally, any control mechanism) is unproblematic. So I rely on some unstated classificatory schemes, mereological principles and concepts of boundary – specifically distinctions between control mechanisms and their parts and the worlds they inhabit. For example, our definition of a thermostat includes the heating component, such as a home furnace, as part of the thermostat, but not the air molecules close to the heating element or the home in which it functions.

I also assume that class instances, such as the specific thermostat in your home, remain exemplars of their class in all the varied circumstances we consider. I exclude broken or malfunctioning thermostats and cases in which some other agency in the world (such as child with a screwdriver) directly interferes with the inner workings. Only when we have a theory of the semantic properties of a class of mechanisms, modulo the worlds in which they function, can we begin to think of more general cases in which the actual mechanism itself is subject to change. So in what follows I fix the definition of control mechanism, keeping it constant, while varying its environment.

8 The metalanguage of causal graphs

To expand Sloman's sketch we need clarity on kinds of 'causal links'. I use causal graphs for this purpose, borrowing from manipulability theories of causation, especially as developed by Pearl (2000) and Woodward (2003). Adopting the formalism of casual graphs is merely a pragmatic choice for this paper. The more important point is that, just as Tarskian semantics requires a 'metalanguage' to describe both the formal system and its model, we also need a metalanguage to describe control mechanisms, their worlds and the causal relations between them.

A causal graph captures the fixed causal structure of some part of a world at a given level of description. It is a directed graph represented by the ordered pair (V,E), where V is a set of vertices that denote variable properties of ontological features of arbitrary type (e.g., ambient temperature, height of mercury column, state of a data structure etc.) and E is a set of pairs denoting directed edges between ver-

tices that represent invariant causal relationships between those features (e.g., temperature is a direct cause of the height of the mercury column etc.) (Pearl, 2000). If we are uncertain about the causal relationship between two features we connect them with an undirected edge that represent covariation (e.g., is a metal's temperature the cause of its thermal expansion or identical with it?) Each variable $v_i \in V$ is some arbitrary function of its direct parents. So we associate a set of functions, $F = \{v_1 = f_1(\cdot), v_2 = f_2(\cdot), \dots\}$, with every causal graph.

9 Control mechanisms

We define the abstract causal structure of a class of control mechanisms in terms the following causal graph.

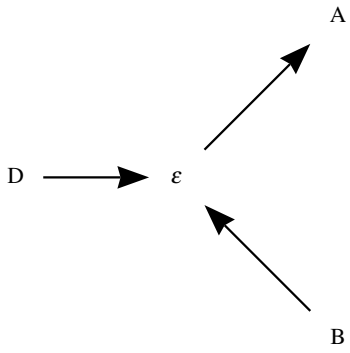


Fig. 1 A causal graph of a control mechanism $\mathscr{K}$.

Definition 1. A *control mechanism* $\mathscr{K}$ is the causal graph

$$(V_{\mathscr{K}}, E_{\mathscr{K}}) = (\{B, A, \varepsilon, D\}, \{(B \to \varepsilon), (D \to \varepsilon), (\varepsilon \to A)\})$$

where B is a belief-like sub-state, A is an action-like sub-state, ε is an error-like sub-state, and D is a desire-like sub-state (see figure 1). The associated function set satisfies

$$\varepsilon = 0 \iff A = \perp ,$$

where $\varepsilon = 0$ means that $\mathscr{K}$'s 'correspondence test' between D and B is satisfied and $A = \perp$ means the 'null action' (i.e. control mechanism $\mathscr{K}$ is not performing an action and therefore A is not the direct cause of any event).

My use of the terms belief-like, desire-like etc. do not imply the information-bearing substates have semantic content (since that has yet to be established). The terms are simply shorthand for their functional role within the control mechanism.

For example, interpret figure 1 as depicting the causal relations in a thermostat. Then variable D represents the thermostat's set point (say $D = 72\ ^{\circ}\mathrm{C}$); variable B is a winding measure of the thermostat's spiral bimetal element; variable A is a measure of the furnace's heat output in Watts; and variable ε, for 'error', measures the discrepancy between B and D, which is the distance between the spur of the spiral coil and the mercury bulb.

Graph $\mathscr{K}$ simply states that B and D are direct causes of ε, and ε is the direct cause of A. Our metalanguage description of control mechanism $\mathscr{K}$ can therefore represent other control mechanisms that share the same casual structure, such as servomechanisms, or (perhaps transient) control mechanisms in more sophisticated information-processing architectures that execute in higher-level virtual machines.

10 Worlds, loops and models

A world is any arbitrarily complex causal graph that doesn't include the individual control mechanism we're considering.

Definition 2. A *world* $\mathscr{W} = (V_{\mathscr{W}}, E_{\mathscr{W}})$ is any causal graph such that $V_{\mathscr{W}} \cap V_{\mathscr{K}} = \{\}$.

A control mechanism $\mathscr{K}$ is open to inputs (which affect its belief-like state B) and can emit outputs (which are caused by its action-like state A). We can therefore form a control loop by connecting $\mathscr{K}$ to a world $\mathscr{W}$ in the 'right way'.

Definition 3. A *loop* $L_{\mathscr{K},\mathscr{W}} = (V_L, E_L)$ is the directed cyclic graph formed by embedding control mechanism $\mathscr{K}$ in world $\mathscr{W}$ such that

(a) $V_L = V_{\mathscr{K}} \cup V_{\mathscr{W}}$,
(b) $E_L = E_{\mathscr{K}} \cup E_{\mathscr{W}} \cup E_*$, where $(A \to v_i) \in E_*$ and $(v_j \to B) \in E_*$ for some $v_i, v_j \in V_{\mathscr{W}}$,
(c) and there is at least one path from A to B.

Assume that all loops are empirically realizable.

Figure 2 depicts a loop formed by embedding $\mathscr{K}$ into a very simple world $\mathscr{W}$ consisting of two variables. Interpret variable E, the direct 'environment', to indicate a feature of the world directly connected to $\mathscr{K}$ and interpret variable W, 'wild causes', to indicate the effect of some other causal agent in the environment. (The functional relations between the variables are not shown, e.g., $E = f_1(A, W)$, $B = f_2(E)$ etc.)

A loop has implicit feedback dynamics in virtue of the functional relations between its variables. Since we make no restriction on the kind of functional relations (they could be arbitrary programs) loops can therefore instantiate arbitrary dynamic systems (whether continuous or discrete, deterministic or probabilistic, symbolic

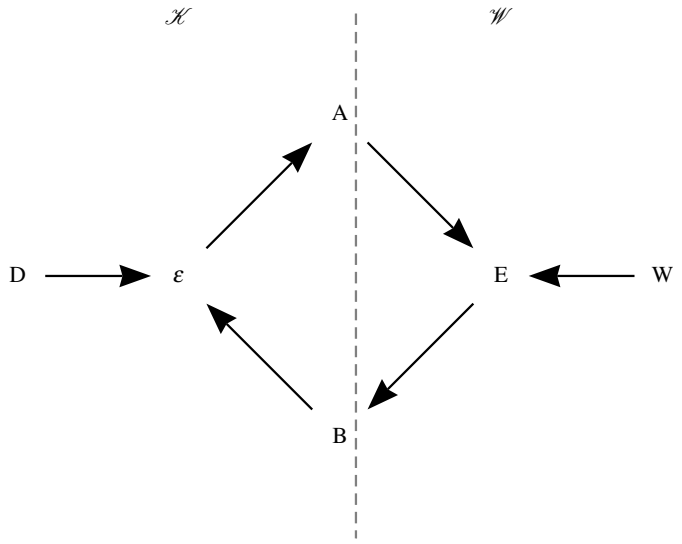


Fig. 2 A loop L formed by embedding control mechanism $\mathscr{K}$ in a world $\mathscr{W} = (\{E,W\},\{(W \rightarrow E)\})$. The dotted line represents the boundary between the control mechanism and its world.

etc.) The only restriction is a fixed causal structure (although finite-sized variable causal structures can be represented by functions that switch sub-graphs in and out).

A model for a control mechanism $\mathscr{K}$ is any world in which it can form a loop.

Definition 4. A *model* is any world $\mathscr{W}$ that can form a loop $L_{\mathscr{K},\mathscr{W}}$ with control mechanism $\mathscr{K}$.

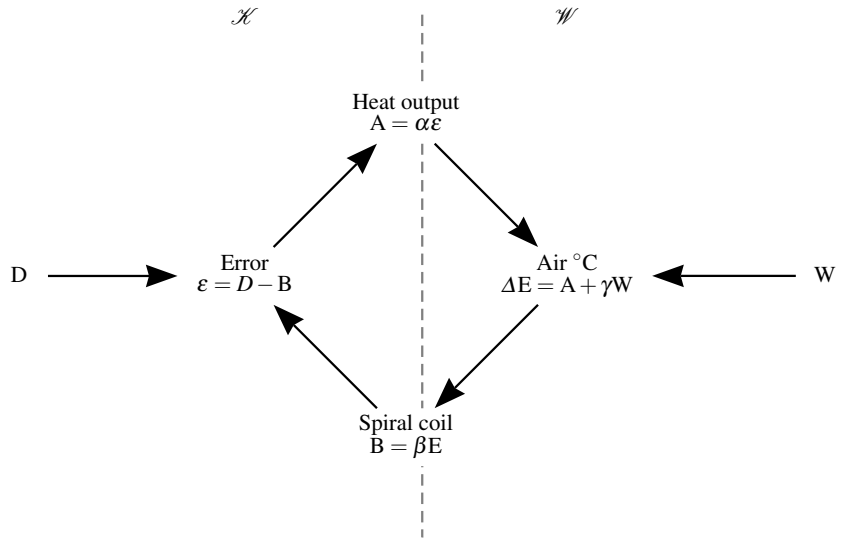
11 Inactivity

For example, figure 3 depicts a thermostat in a room with an open window. Call it model $\mathscr{A}$. The functional relations in this graph are fully specified as differential equations.

In Tarskian semantics the term 'model' is reserved for those domains of interpretation that satisfy a given logic. In loop-closing semantics we reserve the term 'loop-closing model' to denote the subset of models in which control mechanism $\mathscr{K}$ is (i) ultimately inactive and (ii) causally responsible for its inactivity. Let's examine what these conditions mean.

A trajectory in loop $\mathscr{L}_{\mathscr{K},\mathscr{W}}$ is any time-ordered trajectory in state-space generated by the loop's dynamics.

For example, the function set of model $\mathscr{A}$ is $F = \{\varepsilon = D - B, A = \alpha\varepsilon, \frac{\mathrm{dE}}{\mathrm{dt}} = A + \gamma W, B = \beta E\}$ (see figure 3). In this specification D, W, α, γ and β are free



$$\text{(i) } A = \alpha(D - \beta E) \quad \text{(ii) } \frac{dE}{dt} = A + \gamma W$$

Fig. 3 Loop $\mathscr{A}$: A loop formed by a thermostat embedded in a room with a window. D is the thermostat's set point and W is the area of the window open to the outside air. The functional relations between the variables define a simple dynamic feedback system (equations (i) and (ii)).

variables. So model $\mathscr{A}$ in fact defines a family of models. Let's examine two fully specified models from this family.

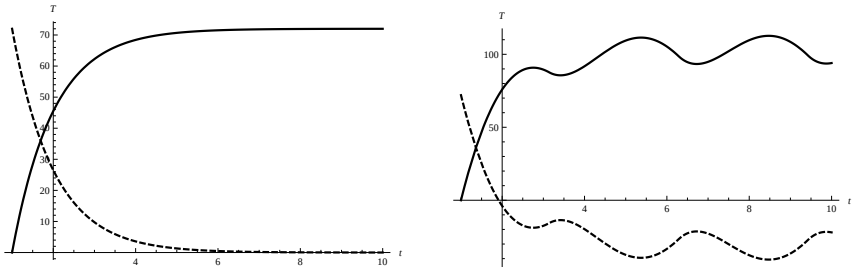
Loop $\mathscr{A}_1$, with $D = 72$, $W = 0$, $\alpha = 1$, $\gamma = 50$, and $\beta = 1$, represents a thermostat with a fixed set-point in a room with a permanently closed window.

Loop $\mathscr{A}_2$, with $D = 72$, $W \approx |\sin t|$, $\alpha = 1$, $\gamma = 50$, and $\beta = 1$, represents a thermostat with a fixed set-point in a room with a window that continually opens and closes (modelled by the $\sin t$ term).

Figure 4 plots both loops given some initial conditions. We immediately see that the thermostat in loop $\mathscr{A}_1$ 'successfully' controls the room temperature whereas the thermostat in loop $\mathscr{A}_2$ fails due to the variable cooling caused by the open window (the thermostat is always playing 'catch up'). The thermostat is 'successful' in the sense that it enters a state of inactivity, or quiescence, in which its action-like state A is not the direct cause of any event.

Definition 5. A control mechanism $\mathscr{K}$ is *inactive* when $\varepsilon = 0$; otherwise, it is *active*.

For example, an active thermostat heats the room (i.e., $A \neq \perp$), which causes its temperature-sensing element to change state. But when the room temperature equals the set-point then heating ceases and (for some non-zero period) the thermostat neither heats or cools (i.e., $A = \perp$).



(a) An ultimately inactive trajectory in loop $\mathscr{A}_1$ (window shut).

(b) A trajectory in loop $\mathscr{A}_2$ (window opening and closing).

Fig. 4 State-space trajectories of thermostat in a room with a window. The solid line represents room temperature and the dotted line represents the thermostat's heat output.

Note that $\varepsilon = 0$ implies $A = \perp$ by the definition of a control mechanism. Inactive control mechanisms are 'in equilibrium' with their environment because the environment corresponds to their desire-like sub-state.

Definition 6. A trajectory is *ultimately inactive* if for some t_0 (i) control mechanism $\mathscr{K}$ is inactive and (ii) for all $t < t_0$ control mechanism $\mathscr{K}$ is active.

The first condition in this definition requires that the control mechanism enter a state of equilibrium or inactivity. The second condition simply excludes degenerate trajectories where $\mathscr{K}$ is never active.

In our simple examples we could use standard Lyapunov stability theory to prove that model $\mathscr{A}_1$ is ultimately inactive (for all d_0) while model $\mathscr{A}_2$ is not. In general, however, such proofs elude us and we must rely on finite, and therefore fallible, observation of the trajectory itself.

A trajectory may be ultimately inactive yet $\mathscr{K}$ need not be the causal agent solely responsible for it. The outcome could be fortuitous, accidental or necessarily depend on the actions of some other agent. For example, consider a variant of model $\mathscr{A}_2$ in which a standalone heating lamp synchronizes its output with the opening and closing of the window. The heat lamp negates the cooling effect of the open window and the loop is then ultimately inactive. But here $\mathscr{K}$ is not 'sole cause' of its own inactivity. Let's turn, therefore, to identifying the conditions in which $\mathscr{K}$ is causally responsible.

12 Output as the actual cause of input

An active control loop, in which $\mathscr{K}$ repeatedly emits actions and senses the consequences, may be initiated by a change in its desire-like or its belief-like sub-states.

In some loops $\mathscr{K}$ is 'in control' because its action A fully determines those features of its environment measured by B. For example, if $\mathscr{K}$ is a thermostat in a

closed room that contains no other sources of heat then, in all likelihood, it will fully control the ambient temperature and hence the winding of its spiral coil. In contrast, if cold air blows through an open window, or a heating engineer directs a hair-dryer towards the thermostat to check it's working, then the thermostat is not fully 'in control'. In these cases the thermostat either partially controls or fails to control its own belief-like sub-state B. Can we make these kinds of distinctions precise?

The concept of 'actual cause' employed in manipulability theories of causation (Woodward, 2003) specifies what it means for $\mathscr{K}$ to be 'in control'. The actual cause is an event 'recognized as responsible for the production of a given outcome in a specific scenario, as in "Socrates drinking hemlock was the actual cause of Socrates death"' (Pearl, 2000, pg. 309). Socrates executed his own death sentence by voluntarily drinking hemlock because he believed the law should be obeyed. His belief was a 'contributing cause' of his death. But it was the ingestion of hemlock that actually killed him.

Pearl's elegant formal theory of causal intervention provides precise definitions of 'actual cause' and related notions (Pearl, 2000). But to maintain a focused exposition I will use a relatively informal definition (borrowing from the formal theory) sufficient for our purposes.

First, we need the idea of an intervention. An 'intervention' is intended to capture 'processes that satisfy whatever conditions must be met in an ideal experiment designed to determine whether X causes Y' (Woodward, 2003, pg. 46).

Definition 7. I is an *intervention variable* for X (the 'causing variable') with respect to Y (the 'effect variable') if and only if (i) I is a direct cause of X, (ii) I is the only direct cause of X, (iii) any directed path from I to Y goes through X, and (iv) if there is a directed path from any other variable V to I that does not go through Y, then any directed path from V to Y goes through X (this definition based on Woodward (2003, pgs. 98-100)).

Consider the causal graph depicted in figure 2. To illustrate the concept let's determine whether ε is an intervention variable for action A with respect to belief B. We can easily see that (i) ε is the direct cause of A, (ii) ε is the only direct cause of A, and (iii) any path from ε to B goes through A. Hence the first three conditions are met. Condition (iv) is more complex. It's designed to exclude cases where a variable V affects both the 'causing variable' X and the 'effect variable' Y but does so via an independent route. If such a V existed then an intervention on X could not be the unique cause of an effect on Y. The causal structure of a control mechanism $\mathscr{K}$ in fact guarantees there cannot be any such V. So condition (iv) is also satisfied and hence ε is an intervention variable for A with respect to B.

Intervention variables support intervention events, or simply 'interventions'.

Definition 8. I taking some value $I = z_i$ is an *intervention* on X with respect to Y if and only if I is an intervention variable for X with respect to Y (Woodward, 2003, pg. 98).

In other words, if I is an intervention variable then I taking a new value z_i is a special kind of event called an intervention. We can now define our target concept, 'actual cause'.

Definition 9. $X = x$ is an *actual cause* of $Y = y$ if and only if there is at least one path R from X to Y for which an intervention on X will change the value of Y given that other causes Z_i of Y, which are not on path R, are fixed by interventions at their actual values (cf. Woodward (2003, pg. 77)).

For $X = x$ to be the actual cause of $Y = y$ then X and Y must be causally connected and other possible extraneous or 'off path' causes of Y are either absent or fixed at constant values. When X is the actual cause of Y then changes in the value of Y are ultimately and uniquely traceable to changes in the value of X.

Note that the definition of 'actual cause' is counterfactual: it does not require that an exogenous intervention on X actually take place. Hence X taking the value x can be the actual cause of Y taking the value y in the context of a self-sustaining loop of causation in which X is itself endogenously caused (such as a loop in which $\mathscr{K}$ is active).

We can now define what it means for a control mechanism to be 'in control'.

Definition 10. A control mechanism $\mathscr{K}$ is *loop-controlling* when its action-like substate $A = a$ is the actual cause of its belief-like substate $B = b$ (via some 'active path' R from A to B).

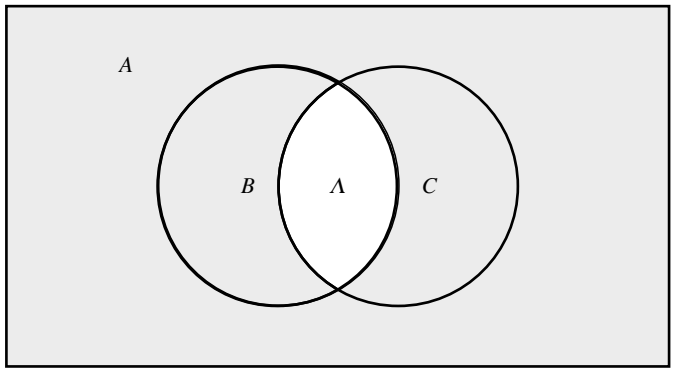
A control mechanism is loop-controlling, therefore, when its output is the actual cause of its input.

Let's return to the variant of model $\mathscr{A}_2$ where a heating lamp contributed to the thermostat's ultimate inactivity. In this case the thermostat is not loop-controlling because the lamp is another cause of the winding or unwinding of its spiral coil. So although the loop is ultimately inactive the control mechanism is not the actual cause of this outcome.

A control mechanism may fail to be loop-controlling even in the absence of 'wild factors' such as W. For example, consider a thermostat that heats water in a container. Set the thermostat's set point to $D = 105$ °C. Water boils at 100 °C at which point its temperature stabilizes during vaporization. At this temperature the thermostat's output is also not the actual cause of its input (since any increase in heating output does not change the winding of its spiral coil).

The definition of loop-controlling handles not just simple cases but scales to arbitrarily complex causal graphs.

Now that we've defined situations when a mechanism controls its own input we can consider complete dynamic trajectories where the control mechanism is fully 'in control'.



$A =$ set of loops for $\mathscr{K}$
$B =$ subset of ultimately inactive loops
$C =$ subset in which $\mathscr{K}$ is loop-controlling
$\Lambda = B \cap C =$ subset in which $\mathscr{K}$ is the actual cause of its own inactivity
$\Lambda_{\mathscr{K}} = \{\mathscr{W} \mid L_{\mathscr{K},\mathscr{W}} \in \Lambda\} =$ the set of loop-closing models.

Fig. 5 Venn diagram representation of the set of loop-closing models.

13 Loop-closing models

To 'close' a loop means to terminate its activity. A control mechanism implicitly defines a set of 'loop-closing models'. A loop-closing model is a world in which it's possible for the control mechanism to be the actual cause of its own inactivity.

Definition 11. A *loop-closing trajectory* is (i) an ultimately inactive trajectory in which (ii) control mechanism $\mathscr{K}$ is loop-controlling when active.

Definition 12. A world $\mathscr{W}$ is a *loop-closing model* for control mechanism $\mathscr{K}$ if there exists at least one loop-closing trajectory in loop $\mathscr{L}_{\mathscr{K},\mathscr{W}}$.

A loop-closing model is like a 'natural' controlled experiment that creates conditions in which $\mathscr{K}$ manifests its casual powers without interference. Controlled experiments are designed 'to get a single mechanism going in isolation and record its effects' since outside of experimental conditions the powers of the mechanism 'will normally be affected by the operations of other mechanisms' such that 'no unique relationship between the [observed] variables or precise description of the mode of operation of the mechanism will be possible' (Bhaskar, 1997, pg. 43). Any countervailing mechanisms that could affect the causal links between $\mathscr{K}$'s outputs and inputs are either absent or inactive – and therefore 'controlled for'. The loop-closing models are experiments that an observer, ignorant of the function and design of the control mechanism, might need to setup in order to identify its essential causal powers.

An analogy might help here. Imagine you find a door key. This key opens some subset of all the locks in the world. The properties of the key – its size, shape and

number of teeth etc. – implicitly define the set of locks it can open. In principle you can enumerate this set by successfully unlocking doors (on condition you control for countervailing and interfering factors, such as rusty locks, broken mechanisms, hidden secondary locks etc.) The loop-closing models, then, are worlds that control mechanism $\mathscr{K}$ 'unlocks' in virtue of its causal powers alone and therefore $\mathscr{K}$, in these circumstances, is the actual cause of its own inactivity. Figure 5 depicts the logical relationship between the set of all possible loops and the special subset of loop-closing models.

Loop-closing models 'define a non-Tarskian model for the internal representations which play a role in percepts, beliefs plans, etc., namely an external environment which can coherently close the feedback loops' (Sloman, 1986b). We've deliberately restricted our discussion of control mechanisms to the relatively simple causal structure $\mathscr{K}$. But in general 'this notion of coherent causal closure will be relative to the system's ability to have precise and detailed goals and beliefs. How specific the mapping is between internal representations and external structures will depend on how rich and varied is the range of percepts, goals and action strategies the system can cope with' (Sloman, 1986b).

Let's turn now to constructing a Tarksian-like mapping between internal representations and the external structures of loop-closing models.

14 The manipulable feature

A control mechanism $\mathscr{K}$ controls some special feature or collection of features in virtue of its causal powers. I call this collection of features the 'manipulable feature'. The set of loop-closing models possess the remarkable property of implicitly specifying what the manipulable feature actually is. Let's now make that explicit.

Definition 13. The *conjunctive feature* of a loop-closing model, $\mathscr{W} = (V_{\mathscr{W}}, E_{\mathscr{W}})$, is the conjunction

$$f_{\mathscr{W}} = f_1 \wedge f_2 \wedge \cdots \wedge f_n,$$

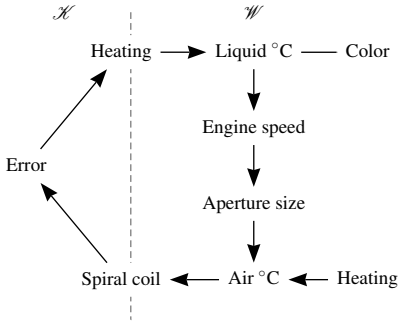
where (i) $f_i \in V_{\mathscr{W}}$ and (ii) there exists at least one loop-closing trajectory in $\mathscr{L}_{\mathscr{K},\mathscr{W}}$ with f_i on its active path.

The conjunctive feature is simply the conjunction of features that connect $\mathscr{K}$'s outputs to its inputs in a loop-closing trajectory.

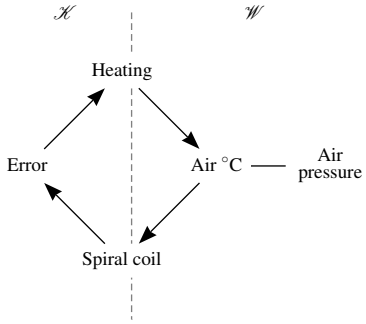
For example, imagine we attach a thermostat to the Rube Goldberg setup we described in section 3. Figure 6(a) depicts the loop's casual graph, $\mathscr{L}_{\mathscr{K},\mathscr{W}_1}$. Assume the functional relations entail that $\mathscr{W}_1$ is loop-closing. The conjunctive feature is then

$$f_{\mathscr{W}_1} = \text{Liquid } ^\circ\text{C} \wedge \text{Color} \wedge \text{Engine} \wedge \text{Aperture} \wedge \text{Air } ^\circ\text{C},$$

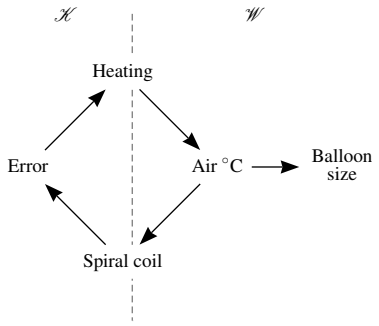
which is simply the conjunction of features on the active path, where these labels denote variable properties (e.g., 'Engine' is shorthand for the rotary speed of a thermokinetic engine).



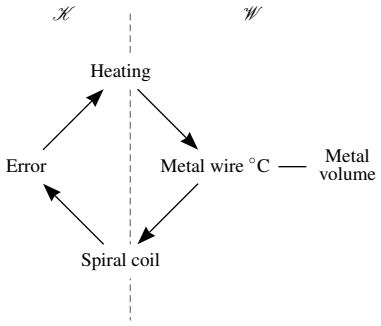
(a) $\mathscr{L}_{\mathscr{K},\mathscr{W}_1}$: Controlling the conjunction of features in a Rube Goldberg loop.



(b) $\mathscr{L}_{\mathscr{K},\mathscr{W}_2}$: Controlling the air temperature in a room.



(c) $\mathscr{L}_{\mathscr{K},\mathscr{W}_3}$: Controlling the air temperature inside a balloon.



(d) $\mathscr{L}_{\mathscr{K},\mathscr{W}_1}$: Controlling the temperature of a metal wire.

Fig. 6 A small set of loop-closing models, $\Lambda_{\mathscr{K}} = \{\mathscr{W}_1, \mathscr{W}_2, \mathscr{W}_3, \mathscr{W}_4\}$, for a thermostat $\mathscr{K}$.

For example, consider a thermostat that controls the temperature of a closed room in which temperature and pressure covary (loop $\mathscr{L}_{\mathscr{K},\mathscr{W}_2}$ in figure 6(b)). The conjunctive feature is

$$f_{\mathscr{W}_2} = \text{Air } ^\circ\text{C} \wedge \text{Air pressure}.$$

Another example: consider a thermostat that controls the air temperature inside a balloon. The balloon expands and contracts and hence temperature and pressure do not covary (loop $\mathscr{L}_{\mathscr{K},\mathscr{W}_3}$ in figure 6(c)). The conjunctive feature in this case is

$$f_{\mathscr{W}_3} = \text{Air } ^\circ\text{C}.$$

Finally, consider a thermostat that controls the temperature of a metal wire soldered to the heating element and the spiral coil (loop $\mathscr{L}_{\mathscr{K},\mathscr{W}_4}$ in figure 6(d)). The conjunctive feature is

$$f_{\mathscr{W}_4} = \text{Metal } ^\circ\text{C} \wedge \text{Metal volume}.$$

The sample of loop-closing models for thermostat $\mathscr{K}$ are then $\Lambda_{\mathscr{K}} = \{\mathscr{W}_1, \mathscr{W}_2, \mathscr{W}_3, \mathscr{W}_4\}$. The models have varied causal structures but share the common property of forming coherent 'causal closures' with $\mathscr{K}$.

Definition 14. The *disjunctive feature*, $f_{\Lambda_{\mathscr{K}}}$, of a set of loop-closing models, $\Lambda_{\mathscr{K}} = \{\mathscr{W}_1, \mathscr{W}_2, \dots\}$, for control mechanism $\mathscr{K}$, is the disjunction

$$f_{\Lambda_{\mathscr{K}}} = f_{\mathscr{W}_1} \vee f_{\mathscr{W}_1} \vee \dots,$$

where each $f_{\mathscr{W}_i}$ denotes the conjunctive feature of model $\mathscr{W}_i$.

In our example, $f_{\Lambda_{\mathscr{K}}}$ is the expression

$$\begin{gathered}(\text{Liquid } ^\circ\text{C} \wedge \text{Color} \wedge \text{Engine} \wedge \text{Aperture} \wedge \text{Air } ^\circ\text{C}) \\ \vee\ (\text{Air } ^\circ\text{C} \wedge \text{Air pressure}) \\ \vee\ (\text{Air } ^\circ\text{C}) \\ \vee\ (\text{Metal } ^\circ\text{C} \wedge \text{Metal volume}),\end{gathered}$$

which is simply the disjunction of all the conjunctive features.

For large sets of loop-closing models the disjunctive feature is highly complex and will typically exhibit redundancies. We want to obtain the minimal set of features necessary for loop closure. (For example, the presence of a thermo-kinetic engine is neither sufficient or necessary for the thermostat's loop to close.) We now state the key organizing concept of this paper.

Definition 15. The *manipulable (or controlled) feature* of a set of loop-closing models $\Lambda_{\mathscr{K}}$ is the minimal disjunctive normal form (MDNF) of the disjunctive feature $f_{\Lambda_{\mathscr{K}}}$.[1]

The manipulable feature is the minimal specification of features that, if present on an active path, imply loop closure is possible. In worlds that lack the manipulable feature the control mechanism cannot be the actual cause of its own inactivity.

We can use the Quine-McCluskey algorithm (McCulskey, 1959) to compute the MDNF expression. For example, the disjunctive feature in our example reduces to the manipulable feature

$$\text{Air } ^\circ\text{C} \vee (\text{Metal } ^\circ\text{C} \wedge \text{Metal volume}).$$

So we can conclude, given this small sample of loop-closing models, that the thermostat's manipulable feature is air temperature *or* the temperature of a metal wire (and its covarying volume). The satisfaction of at least one conjunctive clause of the manipulable feature is sufficient for loop-closure. Simply put, the presence of air or a metal wire between the thermostat's inputs and outputs is sufficient. The

[1] I ignore some technical details here. First, some conjunctive features are necessarily coupled and order-dependent due to the associated functional relations. We can take account of order-dependence but doing so would complicate the exposition without altering the basic argument. Second, the MDNF may not be unique, in which case we have a set of manipulable features.

presence of the other Rube Goldberg transmission mechanisms are not required for loop-closure and therefore excluded from the definition of the manipulable feature.

We cannot conclude that the absence of air or a metal wire implies loop-closure is impossible because we haven't considered the set of all possible loop-closing models. The thermostat might be able achieve loop-closure in worlds we have yet to consider.

Note that, in our example, air pressure is not a manipulable feature because model $\mathscr{W}_3$ – controlling the air temperature inside a balloon – excludes it. In general, we need highly specific worlds (i.e., natural or artificial experimental situations) to 'control for' interfering variables and thereby exclude features.

Additional loop-closing models can further constrain the manipulable feature. Consider a variant of loop-closing model $\mathscr{W}_4$ where we clad the metal wire in ceramic housing to prevent its thermal expansion. In consequence the volume of the metal wire is now constant and does not vary with its temperature. The conjunctive feature of this new loop-closing model is

$$f_{\mathscr{W}'_4} = \text{Metal } {}^\circ\text{C}.$$

Add this feature as a new term to the disjunctive feature and apply the Quine-McCluskey algorithm again. The manipulable feature for the expanded set of loop-closing models is

$$\text{Air } {}^\circ\text{C} \vee \text{Metal } {}^\circ\text{C}.$$

The presence of thermal expansion is therefore unnecessary for loop closure.

The specification of the manipulable feature is of course relative to the choice of a metalanguage. Although we share the same world as a given control mechanism Sloman remarks 'it would be incoherent to try to describe the common underlying reality in neutral terms' (Sloman, 1986a). In consequence, 'Like Tarskian semantics, "loop-closing semantics" leaves meanings indeterminate. For any level of specification at which a loop-closing model can be found, there will be many consistent extensions to lower-levels of causal structure (in the way that modern physics extends the environment known to our ancestors), which remain adequate models in this sense.' (Sloman, 1986b).

For example, the loops depicted in figure 6 happen to be specified at a 'common sense' level of abstraction. The manipulable feature is 'air temperature' or 'metal temperature'. We can extend the metalanguage to include the theory of heat diffusion and re-describe the loops in $\Lambda_{\mathscr{K}}$. The extended metalanguage identifies the common ontological features of the disjunction (i.e., 'jiggling atoms') and collapses the manipulable feature to the natural kind term 'temperature', suitably restricted to the range of values and kinds of functional relations the thermostat can control. (In general, unrestricted 'temperature' features are not manipulable; for example, a domestic thermostat cannot control the temperature of the sun.)

My guess is that, as we expand the set of models to include all possible worlds, then the manipulable feature will converge to the fixed point 'temperature'. This would be the thermometer's *ultimate manipulable feature*. After all, a thermome-

ter has very specific and dedicated input and output channels specially designed to control temperature.

So by a roundabout route we have arrived at the common-sense conclusion that a thermostat in fact controls temperature – but not in virtue of the intentions of its designer, or the uses to which it is normally put, or the reasons for its existence and persistence as an artifact, or the method by which it acquired its causal properties – but *in virtue of the kind of thing it is*. The manipulable feature is an objective property of a control mechanism and its possible control loops.

In general, manipulable features exist without necessarily being fully known, either by an observer, any designer or user of the mechanism, or the control mechanism itself (even if the mechanism is a component of an intelligent agent, such as ourselves).

15 Some reductive definitions

The causal structure of control mechanism $\mathscr{K}$ defines the functional roles of its interrelated substates. We imply those functional roles by calling D 'desire-like', B 'belief-like', A as 'intentional action-like', and ε as an 'error', 'comparison' or 'discrepancy' state. Let's now address how loop-closing theory claims that these substates indeed have the semantic properties implied by their functional-role names.

Recall that Sloman (1997) states that loop-closing theory 'merely constitutes a partial specification of what "belief-like" and "desire-like" *mean*' (my emphasis). We now have the conceptual tools to *define* semantic properties, such as what a belief-like state 'represents', in terms of non-semantic causal relations (from the perspective of an observer using a given metalanguage). So now I reintroduce semantic terms in order to reductively define them.

First, we need one further distinction. Consider again loop $\mathscr{A}_2$ where a thermostat heats a room with a window that continually opens and shuts. $\mathscr{A}_2$ is not loop-closing. However, it would be loop-closing if we intervened and fixed the window in position.

Definition 16. A non loop-closing model $\mathscr{W}$ is a *counterfactual loop-closing model* if it can be transformed into a loop-closing model by a set of interventions that fixes some subset of variables $v \subseteq V_{\mathscr{W}}$ at their initial values.

So $\mathscr{A}_2$, while not loop-closing, is counterfactual loop-closing. In contrast, consider a new Rube Goldberg loop $\mathscr{A}_3$, where we place an inverting contraption between the thermostat's output and input such that increased (resp. decreased) heat output gets inverted to increased (resp. decreased) cooling input. This loop is not loop-closing. Also, the inverting functional relations on the active path cannot be changed by intervention (since the inverting contraption lies on the only active path and we restrict interventions to those that 'hold things constant' rather than modifcations to the causal structure). So $\mathscr{A}_3$ is not counterfactual loop-closing.

In counterfactual loop-closing models the manipulable feature is present – it just so happens that other mechanisms prevent the loop from closing. In models that are neither loop-closing or counterfactual loop-closing the manipulable feature is absent.

Now to the reductive definitions. The vertices (or substates) of $\mathscr{K}$ take a range of values of a given type. For example, in the case of a thermostat, the belief-like state *B* is a scalar winding measure of the spiral coil. We map, in a Tarksian-inspired manner, the state of the information-bearing substates (e.g., the coil is 80% wound) to the *covarying* state of the manipulable feature (e.g., the temperature is 20 °C).

Definition 17. The *semantic content* of belief-like state *B* is the covarying state of the manipulable feature (*B* 'refers' to the manipulable feature). *B* has the *semantic value* 'true' in all loop-closing or counterfactual loop-closing models; and the value 'false' otherwise.

B is 'true' when the manipulable feature is present and 'false' otherwise. Note that we make no mention of whether control is successful or not. Neither do we mention 'ideal' or 'normal' conditions, or higher-order teleological concepts of the proper function of a belief-like state, such as the intended use of an artifact.

Definition 18. The *semantic content* of action-like state *A* is a 'command' to change the covarying state of the manipulable feature (*A* is 'directed at' the manipulable feature). *A* has the *semantic value* 'effective' in loop-closing models; and the value 'ineffective' otherwise.

Definition 19. The *semantic content* of desire-like state *D* is the absence of the manipulable feature in a specific state (*D* 'refers' to an unrealized state-of-affairs). *D* has the *semantic value* 'unsatisfied' when control mechanism $\mathscr{K}$ is active; and the value 'satisfied' otherwise.

And we shall also say that $\mathscr{K}$'s activity is a (simple) kind of 'intentional practice', i.e. goal-directed activity, that is *successful* in loop-closing models, and *unsuccessful* otherwise.

Semantic content is invariant across models since the manipulable feature is a property of the control mechanism. Semantic values, however, vary from model to model, since they are a property of the control mechanism's substates in particular worlds.

To illustrate these definitions let's consider the thermostat again. Loop-closing theory therefore claims that the winding factor of the thermostat's spiral coil represents temperature, its set-point represents a possible temperature, and its heating output is a command to alter the temperature. In loop-closing models, such as example $\mathscr{W}_2$, its belief-like state is true (i.e., the spiral coil veridically represents temperature), its action-like state is effective (i.e., contributes to ultimate success) and its desire-like state is ultimately satisfied.

In loop $\mathscr{A}_2$, in which the window opens and shuts, the thermostat's belief-like state is also true (the spiral coil reliably covaries with the manipulable feature), but its actions are ineffective (i.e., they do not contribute to ultimate success) and its desire-like state is forever unsatisfied.

Let's consider a trickier example. Consider a situation where a thermostat outputs heat into room 1 but measures temperature in room 2, where room 1 and room 2 are thermally isolated from each other. Loop-closing theory states that *B* is false and *A* is ineffective even though *B* is reliably caused by the temperature in room 2 and the temperature in room 1 is reliably caused by *A*. Does this make sense?

Yes. The semantic content of the thermostat's substates is the *manipulable* feature 'temperature'. It's true that rooms 1 and 2 both have temperatures. But the thermostat cannot manipulate (that is, control) the temperature in room 1 nor the temperature in room 2. The manipulable feature of the world necessarily present for control success is absent. Hence it's belief-like state is false and its action-like state ineffective.

In this situation we, as observers, can use the winding of the spiral coil to veridically measure the temperature in room 2. But the substate would possess this content in virtue of our intentional practice. In contrast, when the substate is recruited by the thermostat itself, as part of its own 'intentional' activity, then the spiral coil falsely represents a manipulable temperature with a specific state. In this situation the thermostat's belief-like substate misrepresents the world.

Of course, the thermostat has no concept of temperature. So its semantic content is of the 'nonconceptual' or procedural, rather than declarative, kind. It possesses nothing remotely close to our current theory of heat. Hence it doesn't refer to temperature in the way we do; nonetheless, loop-closing theory claims that thermostats do refer to that same aspect of a 'common underlying reality' that we call 'temperature'.

Definitions 17, 18 and 19 reduce semantic to causal properties. Do they constitute a successful reduction? We need to test the theory to answer this question.

16 Conjunction, disjunction and semantic duality

At the start of this chapter we identified three features of semantic properties – reference, focus and semantic value – that we wanted to reductively explain, at least for simple mechanisms with internal structures that we fully understand. Crude causal theory (IS) attempts to reduce semantic properties to reliable causation between a referent Y and reference X. The 'atom of meaning', in this theory, is lawful covariation, such as the lawful covariation between a thermometer's mercury column and temperature.

The 'crude causal theory' encounters various logical problems: The conjunction problem implies that any 'upstream' feature Y can be the referent of any 'downstream' feature X in a chain of reliable causation. So (IS) cannot explain semantic focus. The disjunction problem implies that (IS) cannot separate the conditions necessary for representation from those necessary for veridical representation. So (IS) cannot explain semantic values, such as the truth and falsity of beliefs or the possibility of misrepresentation. And the problem of semantic duality implies that (IS) is

indeterminate with respect to transparent and opaque content. So (IS) cannot explain reference, that is how a substate represents some thing other than itself.

It's clear that semantic properties cannot be reduced to reliable causation. How does loop-closing theory fare on these issues?

16.1 Semantic focus

The 'atom of meaning' in loop-closing theory is a negative feedback control mechanism. The causal structure of control loops may involve reliable causation (such as heat causing thermal expansion) but 'one-way' cause-and-effect relationships, in themselves, under-determine semantic content. Loop-closing theory avoids the conjunction problem because it identifies semantic content with a specific kind of thing – the manipulable feature – which is necessarily present for control success, but in general may be absent. The manipulable feature may appear anywhere, whether proximate or distal, in a complex Rube Goldberg chain of causation or – even more strikingly – not appear at all.

Semantic focus is ultimately explicable in terms of the limited capabilities of specific feedback control mechanisms with specific internal components and dedicated input and output channels. For instance, a thermostat's manipulable feature is temperature (and not altitude), a cruise control system's manipulable feature is velocity (and not temperature), and an automatic inventory system's manipulable feature is the quantity of a particular good (and not the level of a liquid) etc. Such statements are not merely conventional truisms, nor do they depend on considerations of normal use, but rather derive from the causal properties of the mechanisms themselves, the kinds of loops they can form, and the subset that are loop-closing.

16.2 Semantic value

Loop-closing theory reduces the concept of a 'true belief' to an information-bearing substate that (i) performs a belief-like functional role in a negative feedback control system (i.e., has world-to-mind 'direction of fit', combines with desire-like states to form actions etc.) , and (ii) covaries with the control system's manipulable feature. A 'false belief', in contrast, performs a belief-like functional role and covaries with features of the world, but those features are not the manipulable feature.

Loop-closing theory avoids the disjunction problem because a control mechanism may be active in a loop – and hence instantiate substates that refer to the manipulable feature – without the manipulable feature being present. Representational error is possible because not all reliable causes of a belief-like substate are part of its extension. (I examine an extended example of misrepresentation in section 17.3 below.)

16.3 Representation

The problem of semantic duality prompts the question: why does the winding of a thermostat's spiral coil represent temperature and not, more simply, itself? How does loop-closing theory exclude the 'opaque content' of information-bearing substates?

The problem of semantic duality arises because a given relation of reliable causation can participate in multiple kinds of control loops. For example, the thermal expansion of mercury can participate in the practice of measuring temperature (which selects the transparent content of the mercury column) or the practice of calibrating a thermometer (which selects the opaque content). Our goal-directed use of the variation of the height of the mercury column embeds it within an 'intentional circle' and fixes its content.

Loop-closing theory does not break out of the intentional circle but precisely specifies some simple kinds of causal structures that instantiate intentional circles. A thermostat, in sharp contrast to a thermometer, has causal properties, independent of our practices with respect to it, that instantiate a simple kind of intentional practice that manipulates temperature. In this kind of loop the spiral coil represents temperature (the transparent content) and not itself (the opaque content).

Simple cause-and-effect mechanisms, in contrast, such as thermometers, clocks, rulers, barometers, and accelerometers etc., do not possess semantic content independent of our practice with respect to them. They lack sufficiently rich causal structure to define a manipulable feature. In consequence, they merely possess 'derived intentionality'.

17 Applications of loop-closing theory

Let's now consider some more elaborate applications of loop-closing semantics in order to further illustrate its explanatory properties.

17.1 Memory addressing in CPUs

Sloman often points out, in response to claims that only human or biological minds possess original intentionality, e.g., Searle (1980), that 'there are clearly primitive semantic capabilities in even the simplest computers, for they can use bit patterns to refer to locations in their memories' (Sloman, 2002b). It's true that CPUs use bit patterns in memory to point to other locations in memory. But it's not clear why Sloman considers this capability as a primitive case of original intentionality.

A typical von Neumann CPU architecture, in its essentials, comprises a Turing machine with finite memory and a read/write 'head'. Simplifying somewhat, a special register (some memory in the read/write head), called the Program Counter, points to the next instruction in memory. The CPU fetches the new instruction, by

reading from memory, executes it, and then writes some new state to memory. Hence a CPU performing a fetch/execute cycle is a kind of feedback control system with the local RAM as its environment. In philosophical discussions of Turing Machines the control aspect of computation is often overlooked (see Sloman (2002a) for a corrective).

The CPU executes the machine instruction

```
load reg1,(a)
```

that stores the value at memory address `a` in register `reg1` (where `(a)` denotes the value stored at address `a`). Next the CPU fetches

```
add (reg1),5
```

and executes it. Let's stretch our terminology a little. Interpret '`add (reg1),5`' as a desire-like state to add 5 to the contents of address '`reg1`'. The belief-like state is the contents of address '`reg1`'. The action-like state adds 5 to this value. The `add` instruction is successful if the content of address '`reg1`' is updated. However, if '`reg1`' points beyond addressable memory the CPU throws an exception and the instruction fails.

A full loop-closing analysis would require a detailed examination of the causal structure of the microprocessor's logic circuits. Nonetheless it's clear that 'local addressable memory' is part of the manipulable feature since the instructions fail if the CPU is not linked to RAM.

A CPU manipulates the contents of RAM. On a loop-closing account, therefore, the semantic content of the bit pattern stored at memory address '`a`' is the addressable memory at '`(a)`', a conclusion that machine code programmers would find entirely underwhelming. The only additional insight loop-closing theory adds is that the semantic properties are reducible to the causal properties of the CPU's fetch/execute cycle. So this primitive semantic capability is not derivative on our practice with respect to the computer or our interpretation of its 'symbols'. For example, if the CPU happens to be executing a database program then the bits stored at address '`a`' might refer, for users of the software, to an individual customer. But all this means is that the bit pattern at '`a`' participates in multiple co-existing control loops or 'intentional practices'.

17.2 Dretske's marine bacteria

Dretske (1986) discusses the case of marine bacteria that 'have internal magnets (called magnetosomes) that function like compass needles'. The bacteria use their flagella to propel themselves forward while magnetic forces align their bodies parallel to the earth's magnetic field. In the northern hemisphere they tend toward magnetic north. Bacteria in the southern hemisphere, in contast, have their magnetosomes reversed and tend toward magnetic south. In both hemispheres the magnetic field lines point away from oxygen-rich surface water to oxygen-free sediment at the bottom of the ocean.

Jacob (1997, Ch. 4) asks whether the magnetosome's alignment represents magnetic north or anaerobic conditions. Magnetic north 'carries information' or reliably covaries with the direction of anaerobic conditions. Hence, from the perspective of a 'crude causal theory', the semantic content of the magnetosome's alignment is indeterminate due to the conjunction problem.

Jacob adopts a teleological theory of the 'etiological function' of the magnetosome. An etiological function is an intrinsic causal power of a mechanism that explains its persistence through time. The magnetosome, as a magnetic material, intrinsically indicates magnetic north and this property confers survival advantage that was selected by evolution. Hence the magnetosome represents magnetic north.

Dretske (1986) notes a problem with this conclusion: misrepresentation now seems impossible. If we transplant a southern bacterium to the North Atlantic 'it will destroy itself – swimming upwards (towards magnetic south) into the toxic, oxygen-rich surface water'. Nonetheless the magnetosome, on the etiological account, is a veridical representation of magnetic north, despite the dire consequences for the organism.

Millikan's bio-semantic theory identifies content with the conditions necessary for a mechanism to contribute to the 'proper function' of the whole organism (Millikan, 1984). The proper function is fixed by evolution but, unlike the etiological approach, need not be an intrinsic causal power of the mechanism. We might claim that a magnetosome's proper function is to point the bacterium toward oxygen-free water. So a transplanted bacterium's magnetosome points in the 'wrong direction', misrepresenting the location of oxygen-free water.

My brief mention of these teleo-semantic theories of content cannot do them justice. I mention them only to draw contrast with the loop-closing approach. But, at first blush, it seems mistaken to locate semantic properties in the history of types of things rather than the causal properties of tokens of things. An exact functional copy of a marine bacterium, constructed in some futuristic lab, although lacking an evolutionary history should nonetheless have the same semantic properties as a marine bacterium in the wild (this is the standard 'swampman' objection to teleo-semantic theories (Davidson, 1987); for further objections see Fodor (1990, ch. 3)). As it happens, scientists in fact create artificial magneto-tactic micro-organisms (Kim et al, 2010).

Let's apply loop-closing theory to marine bacteria. First, a caveat: this is a philosophical thought experiment designed to test concepts and not a substitute for a detailed analysis of the biological design of a specific kinds of bacteria (e.g., certain species of marine bacteria navigate using a combination of magneto-taxis and aerotaxis (Bazylinski and Frankel, 2004)).

The first question, from a loop-closing perspective, is whether a control mechanism (or mechanisms) recruits the magnetosome to perform a functional role as an information-bearing substate. Now the orientation of the bacterium is an entirely passive affair – it aligns north just like a compass needle (Chen et al, 2010). The magnetic force on the magnetosome 'biases' the bacterium's direction of motion but the magnetosome does not function either as a belief-like or desire-like state within a feedback control mechanism: the bacterium is simply blown about like

a leaf in the wind. We merely have a 'one-way' cause-and-effect relationship between the mangetic field and the magnetosome. In consequence the orientation of the magnetosome, or the force it applies to the bacterium's body, does not map to any manipulable feature. The magnetosome is therefore like a mercury column in a thermometer rather than a spiral coil in a thermostat. Any semantic content it may have – such as pointing to magentic north or aneorobic conditions – is derivative.

Loop-closing theory states that a magnetosome has no semantic content in virtue of the causal powers of the bacterium. This conclusion directly contradicts teleosemantic theories that, in comparison, look beyond the kind of thing the bacterium is and, in consequence, conflate the magnetosome's semantic properties with its possible adaptive functions on an evolutionary timescale.

17.3 A Fodor machine

Fodor (1990, ch. 3) introduces a helpful thought experiment to illustrate the problem of misrepresentation. You are in a park in the evening. You see a dog before you. It causes the symbol 'dog' to be instantiated in you mind, as part of a belief-like state B, 'there's a dog'. Later a furry, four-legged animal crosses your path, which also causes the instantiation of the symbol 'dog'. But – it turns out – the animal is in fact a cat. You were confused by the dim evening light. In this case, belief B is false.

Fodor explains that according to (IS) this is not a case of misrepresentation. Instead the symbol 'dog' refers to the disjunction *dog* or *cat-in-dim-light*, since the latter also reliably causes the activation of the symbol 'dog'. In fact, the problem is even worse than this, since *cardboard-dog-in-dim-light* and umpteen other 'dog-like' things might trigger your thought 'there's a dog'. How does loop-closing theory avoid the disjunction problem in Fodor's thought experiment?

Fodor's example is pitched at the level of conceptual content, whereas my account of loop-closing theory in this paper is restricted to non-conceptual content. I wish to avoid the complication of the possibility of mismatches between a concept in an explicit ontology and the non-conceptual manipulable features that may be associated with it (a possibility that may explain how higher-order false beliefs can cause control success). For simplicity, therefore, I will again avoid the complexities of the human mind and consider a rudimentary machine.

Imagine a 'Fodor machine', denoted $\mathscr{K}_1$, designed to scare dogs away. A camera generates input and a loud siren generates output. A software classifier maps the camera's input to a boolean B. We, as designers of the machine, intend $B = 1$ to indicate the presence of dogs, and $B = 0$ to indicate their absence. We train the classifier from data using some form of supervised learning and achieve a 5% false-positive rate on the test set. The machine compares B against a constant value $D = 0$. If $B \neq D$ the machine sends the signal $A =$ 'on' to the siren, which starts blaring; if $B = D$ the machine sends the signal $A =$ 'off', which deactivates the siren.

The Fodor machine is a control mechanism with belief-like, desire-like and action-like sub-states. As we hoped, when we test in the wild its loop-closing mod-

els include situations where dogs approach the camera, set off the siren and run away (where we experimentally control for dogs chasing nearby rabbits, cars honking horns and so forth, such that the Fodor machine is the actual cause of its own success).

However, in dim light, cats are a false positive for the classifier. In consequence, some loop-closing models include situations where cats approach the camera and get scared away. We designed $\mathscr{K}_1$ to be a dog-scaring machine but its manipulable feature turns out to be (at least) *dog* or *cat-in-dim-light*.

Loop-closing theory states that $\mathscr{K}_1$'s belief-like and desire-like states refer to this disjunction of features. In consequence if $B = 1$ when cats are in view this is not a case of misrepresentation, even though, from our perspective, the device is not working as intended.

Now we place a cardboard cutout of a dog before the camera. In dim light the cardboard dog is also a false positive for the classifier. The siren starts blaring but, of course, the 'dog' does not move. This world is not loop-closing. Machine $\mathscr{K}_1$ is active – it 'tries' to scare the cardboard dog away – but the manipulable feature is absent, and therefore control success, i.e. a state of inactivity, is not attained. $\mathscr{K}_1$'s manipulable feature does not include *cardboard-dog-in-dim-light*. In consequence if $B = 1$ when the cardboard dog is in view this *is* a case of misrepresentation because B falsely represents the presence of the manipulable feature, viz. *dog* or *cat-in-dim-light*.

The point is this: some mechanisms may be highly dedicated systems with univocal focus; others may successfully control a wider collection of disparate things and therefore possess a manipulable feature expressed as a disjunction of features in a metalanguage. Some semantic indeterminacy should be expected. For example, a necessary condition for cuckoo brood parasitism is that both real chicks and cuckoo chicks are loop-closing models for the host bird's control mechanisms. However, this kind of indeterminacy does not constitute a disjunction problem in loop-closing semantics because error is always possible: a reliable cause of a belief-like substate (e.g., *cardboard-dog-in-dim-light*) need not be part of its extension (e.g., *dog* or *cat-in-dim-light*). As Fodor (1990, pg. 59) remarks 'the least you want of a false token is that it be caused by something that is not in the symbols extension'.

17.4 Ramsay-Whyte success semantics

The loop-closing account of the semantic value of beliefs shares common features with the philosophical theory of Ramsay-Whyte success semantics (Whyte, 1990, 1991), which is 'an heir to the pragmatist tradition' (Blackburn, 2010). Success semantics is pitched at the level of abstraction of the propositional attitudes and – simplifying – states that 'the truth of beliefs explains the success of the actions they cause' (Whyte, 1990). The key idea is that actions succeed if our beliefs represent the world correctly. For example, I believe the animal before me is a dog. This belief may be true or false. I decide to make the animal bark by shouting and running

around it. If the animal really is a dog then, all other things being equal, I succeed; otherwise (say, it's a cat on a dim evening) – I fail.

Crane (1995) summarizes success semantics (SS) as

> (SS) A belief-like state B is a veridical representation of state β if and only if actions A caused by B and desire-like state D would succeed if β obtained.

The relation to loop-closing semantics is clear. Both approaches reduce semantic properties to the 'practical' or pragmatic consequences of belief-like, desire-like and action-like states (i.e. episodes in which an agent attempts to 'control' its environment).

A common criticism of (SS) is that 'success' is a semantic property that it fails to reductively define (e.g., Crane (1995, pg. 189)). (SS) defines 'success' as the satisfaction of a desire, which is the bringing about of the desired state-of-affairs. So what a belief represents is defined in terms of what a desire represents. But this just postpones the explanation of representation. (SS) is therefore a circular, not a reductive, explanation of semantic properties (see Whyte (1991) for counter-arguments that attempt to avoid this conclusion).

Loop-closing theory is different in this respect. It ultimately defines 'success' in terms of a state of inactivity (see definitions 5, 6 and 11). Action-like states, unlike beliefs and desires, are *not* representational states, or at least certainly not when the 'rubber hits the road' and the action is performed. Actions may have imperative semantics and ultimately be effective or ineffective with respect to the beliefs and desires that cause them. But such semantic properties are irrelevant to the determination of whether an action is performed or not. von Wright (1971) puts the matter nicely: 'The connection between an action and its result is intrinsic, logical and not causal (extrinsic). If the result does not materialize, the action simply has not been performed. The result is an essential "part" of the action.' The performance or non-performance of an action is therefore a kind of event that can be defined and identified without mention of semantic properties. 'Success', in loop-closing semantics, is defined reductively in terms of inaction. We'd hope that a causal account of semantic properties would ultimately 'ground' reference in terms of actions – and this is precisely what loop-closing semantics achieves.

18 Conclusion

Sloman's sketch of loop-closing semantics aims to reductively explain a subset of semantic properties in terms of the causal powers of generalized control systems interacting with their environments. In this paper I have developed Sloman's sketch, in particular by specifying a metalanguage to describe the causal structure of loop-closing models and identifying a control mechanism's manipulable feature. The manipulable feature is that subset of the world necessarily present for control success, where 'success' is a control episode in which the mechanism is the actual cause of its

own inactivity. The manipulable feature is the key concept that enables loop-closing semantics to explain semantic reference, focus and value.

Thermostats are mundane artifacts that might not seem worthy of philosophical reflection. But simple control systems encourage us to critically re-examine the (often pre-theoretic) concepts we employ to specify and describe semantic properties, such as 'refer', 'represent', 'believe', 'desire' etc. Sloman (1994a) writes: 'I claim that there are no well-defined concepts that correspond to our normal use of these words. Rather there are many different features to be found in the contexts in which we ordinarily talk about meaning, and different subsets of those features can occur in connection with control states of various kinds of machines.'

Control systems, such as thermostats, are qualitatively different from passive systems, such as thermometers. Thermostats possess semantic properties in virtue of their causal powers; in contrast, thermometers possess semantic properties in virtue of our intentional practices. The causal structure of a simple control system is sufficient, and perhaps necessary, to instantiate semantic properties. In consequence, simple control systems, such as the humble thermostat, constitute a fundamental building block from which to construct a general, naturalized theory of semantics.

Acknowledgements My thanks to Aaron Sloman whose ideas and general approach are of course the main inspiration for this paper. Thanks also to Andrew Trigg, Brian Logan and attendees of the 2011 symposium in honour of Aaron Sloman, held at University of Birmingham, who provided useful feedback on an earlier version of this chapter.

References

Barwise JK, Seligman J (1997) Information flow: the logic of distributed systems. Cambridge University Press, Cambridge

Bazylinski DA, Frankel RB (2004) Magnetosome formation in prokaryotes. Nature Reviews, Microbiology 2:217–230

Bhaskar R (1997) A Realist Theory of Science. Verso Classics, Original edition published by Leeds Books Ltd 1975

Blackburn S (2010) Success semantics, Oxford University Press, Oxford, pp 181–199

Chen L, Bazylinksi DA, Lower BH (2010) Bacteria that synthesize nano-sized compasses to navigate using earth's geomagnetic field. Nature Education Knowledge 1(14)

Crane T (1995) The mechanical mind, 2nd edn. Routledge, London

Davidson D (1987) Knowing one's own mind. Proceedings and Addresses of the American Philosophical Association 60:441–458

Dennett DC (1997) True believers: the intentional strategy and why it works. In: Haugeland J (ed) Mind Design II: Philosophy, Psychology, Artificial Intelligence, Massachusetts Institute of Technology, Massachusetts, Reprint of a 1981 publication

Dretske F (1981) Knowledge and the flow of information. MIT Press, Cambridge, MA

Dretske F (1986) Misrepresentation. In: Bogdan R (ed) Belief, Clarendon Press, Oxford

Feferman AB, Feferman S (2004) Alfred Tarski, life and logic. Cambridge University Press, Cambridge

Fodor J (1989) Psychosemantics: the problem of meaning in the philosophy of mind. MIT Press, Cambridge, MA

Fodor J (1990) A theory of content and other essays. MIT Press, Cambridge, MA

Jacob P (1997) What minds can do. Intentionality in a non-intentional world. Cambridge studies in Philosophy, Cambridge University Press, Cambridge

Kim DH, Cheang UK, Köhidai L, Byun D, Kim MJ (2010) Artificial magnetotactic motion control of Tetrahymena pyriformis using ferromagnetic nanoparticles: A tool for fabrication of microbiorobots. Applied Physics Letters 97:17,302–17,303

McCarthy J (1979) Ascribing mental qualities to machines. In: Philosophical perspectives in Artificial Intelligence, Humanities Press, pp 161–195

McCulskey EL (1959) Minimization of boolean functions. Bell Systems Technical Journal 35:149–175

Millikan RG (1984) Language, thought and other biological categories. MIT Press, Cambridge, MA

Pearl J (2000) Causality, 2nd edn. Cambridge University Press, Cambridge

Searle J (1980) Minds, brains and programs. Behavioral and Brain Sciences (3):417–457

Sloman A (1986a) Reference without causal links. In: European Conference on Artificial Intelligence, pp 191–203

Sloman A (1986b) What sorts of machines can understand the symbols they use? Proceedings of the Aristotelian Society, Supplementary volumes 60:61–95

Sloman A (1994a) Representations as control sub-states, cognitive Science Research Centre, School of Computer Science, University of Birmingham, UK

Sloman A (1994b) Semantics in an intelligent control system. Philosophical Transactions of the Royal Society: Physical Sciences and Engineering 349:43–58

Sloman A (1997) Beyond Turing equivalance. In: Millican P, Clark A (eds) Machines and Thought: the Legacy of Alan Turing, vol 1, Oxford University Press, Oxford

Sloman A (2002a) The irrelevance of turing machines to artificial intelligence. In: Scheutz M (ed) Computationalism: New Directions, MIT Press, pp 87–128

Sloman A (2002b) The mind as a control system. In: Hookway C, Peterson D (eds) Proceedings of the 1992 Royal Institute of Philosophy Conference 'Philosophy and the Cognitive Sciences', Royal Institute of Philosophy, Cambridge University Press, Cambridge

Tarski A (1956) Logic, Semantics, Metamathematics: Papers from 1923 to 1938 by Alfred Tarski. Oxford: At the Clarendon Press, Oxford

von Wright G (1971) Explanation and understanding. Cornell University Press, Ithaca, New York

Whyte JT (1990) Success semantics. Analysis 50(3):149–157

Whyte JT (1991) The normal rewards of success. Analysis 51(2):65–73
Wiener N ([1948] 1975) Cybernetics: or control and communication in the animal and the machine, 2nd edn. The MIT Press, Cambridge, Massachusetts
Woodward J (2003) Making things happen. Oxford University Press, Oxford