Reprinted from the *Working Notes of the Symposium on AI and Limited Rationality,* March 28–30, 1989, pp. 18–22, American Association for Artificial Intelligence Spring Symposium Series, Stanford University, California.

Mental Constitutions and Limited Rationality (Extended Abstract)

Jon Doyle[*]
MIT Laboratory for Computer Science
545 Technology Square
Cambridge, Massachusetts 02139

Abstract

One attractive route to developing a theory of limited rationality is through the notion of rational control of reasoning. Rational control is sometimes viewed in terms of rational allocation of limited computational resources in drawing conclusions, with ideal rationality just the infinite-resource case of resource-bounded inference. But some important sorts of limitations on rationality due to rational control of reasoning subsist independently of any limitations on resources. One of these is the notion of the agent's *constitution*, a set of restrictions placed on the permissible states of the agent by the agent or by its designer. While the theory of ideal rationality permits agents to have any consistent and complete set of beliefs and preferences (probabilities and utilities), agents limited by constitutions may forbid some of those sets. We motivate and formalize some elements of a theory of mental constitutions.

Introduction: Limits to rationality

Artificial intelligence is concerned with agents whose rationality is limited in comparision with the agents presumed by decision theory. In decision theory, a rational agent bases its actions on its beliefs about the relative likelihoods of various events and contingencies and its preferences among these. It requires that these beliefs and preferences are consistent and complete enough to determine probability and utility measures. The limitations accepted as unavoidable in artificial intelligence include having to work with less information (more incomplete, possibly inconsistent sets of beliefs and preferences) and with bounds on the available computational resources. Such limitations have led to theories of resource-bounded reasoning in which only some of the consequences of basic beliefs are computed when choosing actions. It is natural to view these theories as cases of rational control of reasoning, in which the agent rationally allocates its resources in each step or episode of reasoning.

One conception of rational control of reasoning views the task of reasoning as that of computing a sufficiently complete and consistent set of beliefs and preferences from those manifest in memory. In this case, the ideally rational agent is just an agent with unbounded resources available to compute an entire complete and consistent set of attitudes. Limited agents, in contrast, will compute sets of attitudes which are complete in only some ways, and which may contain some sorts of implicit inconsistencies. In some cases, we may view these partially complete and consistent sets of attitudes as sets complete and consistent with respect to a weaker than ordinary logic (e.g., as does Konolige [1985]).

But this conception of resource-bounded reasoning does not provide an entirely adequate theory of limited rationality by itself, as some important sorts of limitations on rationality subsist independently of any limitations on resources. Some of these limitations, such as the apparent impossibility of reconciling multiple preference criteria in default reasoning shown by Doyle and Wellman [1989], stem from our wish to facilitate the construction of complex individual agents by decomposing them into sets of simpler "mental agents" or faculties. Another sort of decomposition is that involved in viewing the agent's mental states as sets of attitudes (e.g., beliefs, preferences, intentions). This decomposition leads to more resource-independent limitations, namely restrictions placed on the permissible states of the agent by the agent or by its designer. For example, states might be forbidden to contain certain consistent beliefs or combinations of beliefs, or might be required to contain other beliefs. Such restrictions do not violate any axiom of decision theory by themselves, since decision theory does not require that all consistent and complete sets of beliefs and preferences be acceptable. But these restrictions would seem to violate the spirit of decision theory, whose presumption that all consistent and complete sets of attitudes be possible is clearly visible in much of

[*]Research supported by National Institutes of Health Grant No. R01 LM04493 from the National Library of Medicine.

the literature, especially in group decision theory. We call the restrictions used in defining the agent's permissible or legal states the agent's *constitution.* (Properly speaking, we focus on *mental* constitutions, and ignore related issues posed by the agent's physical makeup, such as its appendages, sensors, etc.)

Due to space limitations, this extended abstract can only present a few elements of the theory of mental constitutions. We first describe the basic concepts of constitutional reasoning informally, drawing on familiar concepts of rational and deliberate action to motivate the various notions. We then present a mathematical formalization of this conception of constitutional reasoning which abstracts away some of the inessential details of the motivation in order to achieve a theory covering a variety of representational systems. A more complete treatment of these ideas, including applications of the formalism to describing some aspects of AI architectures, may be found in [Doyle, 1988] (which improves on earlier treatments in [Doyle, 1983a, Doyle, 1983b]). The full paper will contain a comprehensive treatment of the theory.

Rationality and constitutions

Constitutions interact with the notion of rationality in several ways. The primary way is that constitutions restrict the range of possible choices available to the agent. Thus if rational reasoning is defined to be rational selection of a new set of attitudes (rational with respect to the current beliefs and preferences), then agents with nontrivial constitutions may not be able to reason as rationally as agents with no restrictions on their attitudes, since they must choose successive states from among those allowed by their constitution. Because of this, constitutions play a central role in enforcing a notion of "automatic" or "background" inference, inference performed whether or not it is worthwhile in the current situation. For example, if the agent is required to hold one belief if and only if another belief is absent, then adding the absent belief means removing the present belief, even though the intended step of reasoning did not involve removing the present belief. We call this component of reasoning *constitutional* reasoning since its performance is supposed to be part of the constitution or makeup of the agent, carried out independently of its considered activities.

Given that constitutions may limit the rationality of an agent, it is natural to ask why one would ever seek to impose such limitations. The answer is that restrictions on reasoning abilities can have advantages as well as disadvantages, such as when they steer the agent away from activities expected to be useless and toward activities expected to be useful. This is important since there are often discrepancies between what is rational in the short run and what is rational in the long run. It is often costly to repeatedly discover and consider long term consequences when choosing immediate actions. Constitutions permit the agent to avoid the costs of maintaining long-term rationality by simply avoiding consideration of alternatives that dominate in the near-term but have bad long-term consequences.

A classic example is that of Ulysses, who had himself bound to his ship's mast so that he could hear the Sirens in safety (see [Elster, 1979]). He knew that to hear them was to be drawn irresistably to them, regardless of the ultimately fatal consequences, but that bondage would rule out the possibility of acting to reach them. Computational examples are also common in artificial intelligence. One is the use of reasons or justifications in reason maintenance. These force certain conclusions to be held or avoided, without requiring any sort of conscious consideration by the agent. This is reasonable when the consequences of holding or avoiding these conclusions are sufficiently obscure to the short-term, focused reasoner. In such cases, posing again the decision of whether to draw or avoid these conclusions would allow the agent to overlook the long-term benefits of basing its actions on these principles. This situation is very common and important, since as with humans, it is much easier to act on the basis of what one has learned from experience than to recall why one learned what one learned or what experiences lead one to learn particular things.

Thus a constitution acts as a tool or resource which might be exploited to improve the overall rationality of the agent as well as a source of restrictions on the rationality of an agent. But we need not view constitutions solely as constructs of the reasoner's designer. As the example of reason maintenance suggests, we may also design agents which have the power to specify or modify their own constitutions, making changes they choose rationally to serve their own purposes. In agents designed in this way, rational control of reasoning involves rational emendation of the constitution of the reasoner as well as rational choice of steps of reasoning.

External and internal constitutions

The simplest way to formalize constitutions is to take the *external* (or eternal, or designer's) perspective, in which a constitution is just an axiomatization of the agent's state space, formulated in a language over sets of attitudes. From this perspective, the agent's states resembles database states, and constitutions are essentially the same as sets of integrity constraints, as the notion appears in database theory (see especially Reiter [1988], who views integrity constraints as statements about the database contents).

External constitutions, however, need not come with any distinguished structure, as they can be any set of axioms we desire. Viewing constitutions as controllable resources means it is more interesting to take the *internal* (or temporal, or agent's) perspective, and view a constitution as a set of constraints on states and transitions changable at will in whole or in part. Where the external perspective defines a single set of legal states constituting all the states legal at any time in some history, the internal perspective views the constitution as a time-varying structure, with different sets of legal states at different times. In the external perspective, the agent's constitution is something external to and unchangable by the agent. In the internal perspective, the constitution is internal to and changable by the agent.

To reflect this internal view as time-varying structures, we formalize the notion of constitution so that each constitution has two (possibly vacuous) parts. The first is the fixed part, the most important form of which is called a *constitutive logic*, that is, a logic that specifies the minimal consistency and closure properties of states. The second part is the variable part, called the set of *laws of thought* or *constitutive intentions*, that is, the rules for self-regulation that the agent may adopt or abandon. The *legal* or constitutionally permissible states of the agent are then those states closed and consistent with respect to the constitutive logic and legal according to each of the laws of thought they contain.

The external and internal perspectives can be assimilated by representing internal constitutions as distinguished subsets of the agent's attitudes, and having the external constitution set forth the meanings of all possible laws of thought. One example of such an external form of internal constitutions is given by Minsky [1988], who studies internal constitutions for complex programming systems under the name of "law-governed systems."

Constitutions and framings

To formalize internal constitutions, we view each possible history of the agent as a discrete sequence

$$\ldots, S_{t-1},\ S_t,\ S_{t+1}, \ldots$$

of internal states. We write the set of all possible instantaneous states of the agent as $\mathcal{I}$. In this setting, reasoning is change of view; that is, each step of reasoning may be viewed as a change of state $S \mapsto S'$ for some $S, S' \in \mathcal{I}$ (though not every change of state need be considered a step of reasoning).

In the most basic sense, the agent's constitution is simply its state space $\mathcal{I}$ (or better, its set of possible histories). These states represent the agent's makeup, and the constitution restricts the makeup of the agent to these states and forbids it from taking on forms outside of $\mathcal{I}$. But this sense of constitution is a weak one. The interesting questions begin to arise only when we seek to interpret histories as histories of a rational agent, and to interpret states in terms of the mental attitudes held by the agent. Such interpretations we call (rational or attitudinal) *framings*.

Now it may be possible to frame any set of possible histories as the histories of a sufficiently complex rational agent. (Indeed, the social and economic analyses of economists are full of such rationalizations of observed behaviors.) Since our focus here is on the structure of designed constitutions rather than their identification through observation, we will avoid most questions of interpretation by assuming a standard attitudinal framing of mental states. Specifically, we view each instantaneous state as a set of beliefs (including comparisons of relative likelihood), desires (including preferences), and intentions. This assumption is in accord with standard AI practice, which employs a database of sentences or sententially-interpreted structures to represent states. We let $\mathcal{D}$ stand for the set of all possible mental attitudes, and frame each state $S \in \mathcal{I}$ as a set of mental attitudes $S \subseteq \mathcal{D}$. Thus if $\mathbf{P}\mathcal{D}$ is the powerset (set of all subsets) of $\mathcal{D}$, then according to this reading of states $\mathcal{I} \subseteq \mathbf{P}\mathcal{D}$. (Strictly speaking, this notation confuses states with their attitudinal interpretations, but that will not matter here.) Let $\mathbf{B}$, $\mathbf{D}$, and $\mathbf{I}$ denote the sets (respectively) of all beliefs, desires, and intentions in $\mathcal{D}$.

Every framing of an agent implicitly identifies a measure of constitutional reasoning since making a set of deliberately chosen or intended changes in a state's attitudes may yield a set of attitudes outside the chosen state space. That is, if $\mathcal{I} \neq \mathbf{P}\mathcal{D}$, the modified set of attitudes may not be in $\mathcal{I}$, and to accommodate the intended changes some additional, unintended changes must be made to yield a state in $\mathcal{I}$.

Constitutive logics

One of the most fundamental sorts of restrictions that constitutions may place on the agent's attitudes are requirements for certain forms of closure and consistency, that is, that the agent's sets of attitudes form "theories" with respect to some regular logic of attitudes. In decision theory, for example, full logical consistency and closure is required of an ideally rational agent's beliefs and preferences. In artificial intelligence, more limited forms of consistency and closure (i.e., weaker logics) are the rule, such as the closure and consistency enforced on concept descriptions by automatic taxonomic classifiers in languages like NIKL [Vilain, 1985], or the closure under resolution and consistency with respect to explicitly noted contradictions enforced by an ATMS [de Kleer, 1986].

If we view the state space as a data structure, then logics of attitudes provide "upper bounds" on the state space (to be restricted further by laws of thought), or alternatively, provide lower bounds on the internal structure required of individual states. In fact, the theory of data structures offers an elegant and universal way of describing computable data structures as logical theories, in which the (partial) elements of a data structure correspond exactly to the deductively closed and consistent sets in the logic of the data structure. This is done using abstract (language-independent) logics called *information systems*. Following Scott [1982], an information system Σ is defined by three things: a set $\mathcal{D}$ of "finite" or initial data objects, a set $\mathcal{C}$ of finite "consistent" subsets of $\mathcal{D}$, and an "entailment" relation $\vdash$ on $\mathcal{C} \times \mathcal{D}$. These notions define a data type or *domain* by viewing each individual data object as a "proposition" about domain elements, and each set of data objects as a partial description of some domain element, with bigger sets representing better descriptions. When descriptions contain enough "propositions," the sets of data objects characterize (possibly partial) domain elements, and so we may identify the elements of the domain with these sets of data objects. Each domain is characterized by formal notions of consistency and entailment of sets of domain elements, with the elements of the domain being the subsets of $\mathcal{D}$ that are consistent (that is, each of whose finite subsets is in $\mathcal{C}$) and closed under entailment (that is, which contain Y whenever $X \vdash Y$ and they contain X). The set of consistent, closed subsets of $\mathcal{D}$ is written $|\Sigma|$.

Thus to specify constitution for $\mathcal{I}$, we first provide an information system Σ over the framing $\mathcal{D}$ such that each state is closed and consistent with respect to Σ. This means that each state of the agent is an element of the domain defined by Σ, so that $\mathcal{I} \subseteq |\Sigma|$.

Constitutive intentions and satisfying states

It is usually not possible to view all constitutional restrictions on states as following from a logic of attitudes, since widely used nonmonotonic constructs like defaults cannot be captured in what we usually think of as a logic. In some cases nonlogical restrictions might be naturally made part of the fixed constitution, as global conditions on states which refine the constitutive logic. One example might be the rationality of the set of constructive attitudes contained in a state with respect to the state's manifest attitudes [Doyle, 1989]. But many sorts of nonlogical restrictions, such as individual defaults or reason maintenance justifications, are naturally viewed as locally applicable laws of thought or constitutive intentions. Constitutive intentions are intentions strictly about the agent's cognitive structure as opposed to intentions about the agent's environment or the agent's relation to it. More specifically, we mean them to be rules about the agent's mental structure that are always and immediately followed.

To formalize the restrictions expressed by constitutive intentions, we must identify constitutive intentions in states and say what it means for a state to satisfy them. To begin, we assume that each state contains a single set of currently held laws that the current state is required to satisfy. We write $\mathbf{I}^\star \subseteq \mathbf{I}$ to mean the set of all possible constitutive intentions and $\mathbf{I}^\star(S)$ to denote the set of constitutive intentions in state S.

Next, we view attitudes as attitudes towards propositions, and view propositions as sets of possible worlds, where each possible world decomposes into a state of the agent and a state of its environment. Formally, to complement the set $\mathcal{I}$ of possible internal states of the agent we let $\mathcal{E}$ be the set of possible states of its environment, and mildly abusing the notation write $\mathcal{W} \subseteq \mathcal{I} \times \mathcal{E}$ for the set of possible worlds. Like $\mathcal{I}$, the sets $\mathcal{E}$ and $\mathcal{W}$ are givens of the theory. Each subset of $\mathcal{W}$ is a proposition, and we write $\mathcal{P} = \mathbf{P}\mathcal{W}$ to mean the set of all propositions.

Since in rational control of reasoning we are concerned with the agent's reasoning about itself, the internal portions of propositions will be of more interest than full propositions. We call subsets of $\mathcal{I}$ *internal* propositions, and subsets of $\mathcal{E}$ *external* propositions. If $P \subseteq \mathcal{W}$ is a full proposition, we say that

$$\mathbf{i}(P) = \{S \in \mathcal{I} \mid \exists E \in \mathcal{E} \quad (S, E) \in P\}$$

and

$$\mathbf{e}(P) = \{E \in \mathcal{E} \mid \exists S \in \mathcal{I} \quad (S, E) \in P\}$$

are respectively the internal and external projections of P. We also call these the internal and external propositions determined by P. Propositions purely about the agent's own state satisfy the condition $P = \mathbf{i}(P) \times \mathcal{E}$. Propositions purely about the agent's environment satisfy $P = \mathcal{I} \times \mathbf{e}(P)$.

We write $\iota : \mathbf{I}^\star \to \mathcal{P}$ to indicate a meaning function ι for constitutive intentions. That is, if $x \in \mathbf{I}^\star$, the meaning of x is that the agent intends to act to make its world $W = (S, E)$ be one of the worlds in $\iota(x)$. We assume that constitutive intentions are purely about internal states. This means that the environmental portion of the proposition $\iota(x)$ is satisfied by any world, so the intention reduces to the condition that $S \in \mathbf{i}(\iota(x))$. For example, in viewing a reason maintenance justification $\langle \mathrm{IN}(A), \mathrm{OUT}(B), c \rangle$ (where $A, B \subseteq \mathcal{D}$ and $c \in \mathcal{D}$) as a constitutive intention, we assign an internal meaning of

$$\{S \subseteq \mathcal{D} \mid (A \subseteq S \wedge B \cap S = \emptyset) \supset c \in S\}$$

(see [Doyle, 1983b]).

Finally, we say that a set $X \subseteq \mathcal{D}$ is *satisfying* just in case it satisfies each of the constitutive intentions it contains, that is, if $X \in \mathbf{i}(\iota(x))$ for every $x \in \mathbf{I}^\star(X)$. Note that bigger sets may contain more constitutive intentions, and so be harder to satisfy. Our assumption, then, is that every legal internal state of the agent is satisfying. In this way, legal states exhibit something of Rawls' [1971] notion of reflective equilibrium, agreement between the agent's principles and attitudes.

Alternative representations

Putting constitutive logics and intentions together, we have that the legal states of the agent are the closed, consistent, satisfying states. More generally, we can use similar ideas to represent constitutions for agents with any sort of decompositions of states (not necessarily just decompositions into attitudes) by treating all elements of states as possessing constitutive import. That is, we recast the meaning function ι as a function $\iota : \mathcal{D} \rightarrow \mathcal{P}$ giving the constitutive import of each element of states, and require that legal states satisfy the meanings of each of their elements (i.e., $X \in \mathbf{i}(\iota(x))$ for every $x \in X$). This enlargment of domain is innocuous, since we can always give an element x without true constitutive import the meaning $\iota(x) = \mathcal{P}$, which is a vacuous restriction satisfied by every possible state. This reformulation is particularly appropriate when formalizing apparently non-attitudinal representations like reason maintenance justifications.

The abstract notion of constitutive meanings is sufficiently general that we may represent the same state space via several different constitutions. For example, one can make the constitutional meanings carry the burden of the constitutive logic as long as the empty set is a legal state, simply by redefining the meanings of each $x \in \mathcal{D}$ to be $\iota' = \iota(x) \cap (|\Sigma| \times \mathcal{E})$. But the intent is to use individual laws to express modular or local restrictions on states. Similarly, one can attempt to push the restrictions of all legislation into the constitutive logic. This cannot work if the legislation includes nonmonotonic conditions on states, but some sorts of laws of thought, such as monotonic justifications and reason maintenance "nogoods," can be directly translated into information system structures.

Conclusion

Rational control of reasoning is usually conceived of as making repetitive choices about pursuing and abandoning different inferential paths. This conception is important, and is sometimes taken to exhaust the subject of limited rationality. But proper as it is for understanding some sorts of limits to rationality, it neglects or obscures the role of habitual behavior in shaping the limits to rationality that vary only slowly with changing levels of strictly computational resources. Habits are not useful in every area of activity, but they are central to many elements of human thought and action, and are widely exploited in various guises in artificial intelligence systems. The different purposes of different habits and the different computational properties of different representations and mechanizations mean that habits will take many superficially different forms in AI systems. But their meanings all appear to take the same underlying forms, of which the two most important categories are the notions of constitutive intentions and constitutive logics.

References

[de Kleer, 1986] Johan de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28:127–162, 1986.

[Doyle, 1983a] Jon Doyle. Admissible state semantics for representational systems. *IEEE Computer*, 16(10):119–123, 1983.

[Doyle, 1983b] Jon Doyle. Some theories of reasoned assumptions: An essay in rational psychology. Technical Report 83-125, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1983.

[Doyle, 1988] Jon Doyle. Artificial intelligence and rational self-government. Technical Report CS-88-124, Carnegie-Mellon University Computer Science Department, 1988.

[Doyle, 1989] Jon Doyle. Constructive belief and rational representation. *Computational Intelligence*, 5(1):1–11, February 1989.

[Doyle and Wellman, 1989] Jon Doyle and Michael P. Wellman. Impediments to universal preference-based default theories. In Ronald J. Brachman, Hector J. Levesque, and Raymond Reiter, editors, *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning (KR'89)*, pages 94–102, San Mateo, CA, May 1989. Morgan Kaufmann.

[Elster, 1979] Jon Elster. *Ulysses and the Sirens: Studies in Rationality and Irrationality.* Cambridge Univerisity Press, Cambridge, 1979.

[Konolige, 1985] Kurt Konolige. Belief and incompleteness. In J. R. Hobbs and R. C. Moore, editors, *Formal Theories of the Common-Sense World*, pages 359–403. Ablex, Norwood, 1985.

[Minsky, 1988] Naftaly H. Minsky. Law-governed systems. Technical report, Rutgers University, Computer Science Department, New Brunswick, 1988.

[Rawls, 1971] John Rawls. *A Theory of Justice.* Harvard University Press, Cambridge, 1971.

[Reiter, 1988] Raymond Reiter. On integrity constraints. In Moshe Y. Vardi, editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, pages 97–111, Los Altos, 1988. Morgan Kaufmann.

[Scott, 1982] Dana S. Scott. Domains for denotational semantics. In M. Nielsen and E. M. Schmidt, editors, *Automata, Languages, and Programming: Ninth Colloquium*, volume 140 of *Lecture Notes in Computer Science*, pages 577–613, Berlin, 1982. Springer-Verlag.

[Vilain, 1985] Marc B. Vilain. The restricted language architecture of a hybrid representation system. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 547–551, 1985.