

Comparing Preferences

Mauro Rossi*

mauro.rossi@umontreal.ca

ABSTRACT

The orthodox view in economics is that interpersonal comparisons (ICs) of preferences present insurmountable epistemic difficulties and, thereby, have no scientific legitimacy. A recent line of thought argues against this position by investigating how ordinary people make ICs of preferences. The underlying idea is that the problem of ICs can be solved if we can find scientific evidence showing that ICs can be reliably made in everyday life. In this paper, I provide an assessment of this strategy. I consider four arguments attempting to show that the conditions for having reliable ICs can be satisfied. I argue that all these arguments fail and reject this strategy as unsuccessful.

The susceptibility of one mind may, for what we know, be a thousand times greater than that of another. But, provided that the susceptibility was different in a like ratio in all directions, we should never be able to discover the difference. Every mind is thus inscrutable to every other mind, and no common denominator of feeling seems to be possible.¹

1. INTRODUCTION

It is commonplace that, in everyday life, we compare preferences belonging to different people with respect to their intensity. We typically make such comparisons with relative ease. Moreover, we often do not find inter-personal comparisons of preferences more difficult than intra-personal comparisons, that is, comparisons involving our own preferences.² Things change as soon as we consider the matter from a theoretical point of view. Several authors claim that interpersonal comparisons (ICs henceforth) of preferences are either impossible,³ or meaningless⁴ or, at least, that they are not factual claims, but rather normative statements.⁵ The main reason for theoretical scepticism is that, while the empirical evidence is sufficient for establishing comparisons about the intensity of a single individual's preferences, it appears to

* Postdoctoral Fellow, Centre de Recherche en Éthique de l'Université de Montréal (CRÉUM).

¹ S. Jevons, *Theory of Political Economy*, 4th edition, Macmillan, London 1911 (1871, 1st ed.), p. 14.

² See D. Davidson, *Judging interpersonal interests*, in *Foundations of social choice theory*, ed. by J. Elster – A. Hylland, Cambridge University Press, Cambridge 1986, pp. 195-211, reprinted as *Interpersonal Comparisons of Values*, in D. Davidson *Problems of Rationality*, Oxford University Press, Oxford 2004, p. 59.

³ See S. Jevons, *op. cit.*, p. 14. For an early reaction against the impossibility claim, see I.D.M. Little, *A Critique of Welfare Economics*, 2nd ed., Clarendon Press, Oxford 1957 (1950, 1st ed.), Chapter IV.

⁴ See K. Arrow, *Social Choice and Individual Values*, 2nd ed., Wiley, New York 1963 (1951, 1st ed.), p. 9.

⁵ See L. Robbins, *An Essay on the Nature and Significance of Economic Science*, Macmillan, London 1932, p. 139.



be insufficient for establishing comparisons about the intensity of different individuals' preferences. In turn, this has led various areas of scientific research – most notably, economics – to dismiss ICs of preferences as scientifically illegitimate.

This rejection has not been without repercussions. It is true that economic theory does not need ICs of preferences to explain the behaviour of the main economic agents, i.e. consumers and firms. On the other hand, if ICs of preference strengths are not allowed, welfare economics is unable to settle distributive conflicts by considering the relative importance that different individuals attach to competing states of affairs. In social choice theory, it is impossible to aggregate individual preferences in order to obtain a social ranking of alternative states of affairs, which satisfies, amongst the others, a condition of non-dictatorship. Even worse, if ICs turn out to be impossible, the meaningfulness of various ethical doctrines e.g. preference utilitarianism and at least some versions of objective list theories of well-being, is entirely compromised. Finally, the rejection of ICs has a considerable impact on more applied spheres, such as health care and policy making, which require the use of normatively significant measures indicating how different people's preferences compare in terms of strength.

In the course of the years, the constraints posed by the rejection of ICs have started to appear intolerable. As a consequence, several strategies have been explored to give ICs new scientific legitimacy. In this paper, I want to examine one of them in particular. Broadly speaking, the main idea underlying this strategy is that, by investigating how ordinary people make ICs of preferences in everyday life, it is possible to find a solution to the main theoretical problems concerning ICs. Consider the common view of our comparative practice. It is generally held that ICs of preferences are based on the ascription of preferences with specific content and intensity to other people and to ourselves. If this picture is correct, the suggestion is that the solution to the problem of ICs should be based on the analysis of the conditions that ought to be satisfied in order for us to make reliable ascription of preference strengths to other people and to ourselves.

This approach invites us to examine two different kinds of question: the question of mental ascription, that is, the question of how ordinary people assigns mental terms to other people and to themselves; and the question of the meaning of mental states, that is, the question of what ordinary people mean when they employ mental terms. Broadly speaking, there are two main theories of mental ascription: Theory Theory (TT) and Simulation Theory (ST). According to the former, ordinary people ascribe mental states by means of a 'Theory of Mind' that they, more or less tacitly, possess. According to the latter, ordinary people ascribe mental states to others by trying to simulate, or replicate, their mental activity. On the other hand, there are two main theories of the meaning of mental states in contemporary philosophy of mind: (commonsense) functionalism and experientialism. According to functionalism, the meaning of a mental state is given by the set of causal laws in which such a mental state figures and which relate it to inputs, other mental states and behavioural outputs. Instead, according to experientialism, the meaning of a mental state is given by the more or less conscious experiences that the subject has of it.

Given that the problem of comparing different people's preferences concerns a specific type of mental states (i.e. preferences) and one of their properties (i.e. strength), one would expect the existence of a large literature in philosophy of mind connecting the problem of ICs to these fields of research. Instead, and quite surprisingly, philosophers of mind have almost completely ignored the issue of ICs. Alvin Goldman constitutes the only significant exception. Indeed, in his "Simulation and Interpersonal Utility", Goldman attempts to bring the problem



of ICs in line with current debates in philosophy of mind and epistemology.⁶ For the purpose at stake, Goldman's approach has two limitations: it focuses mainly on ICs of happiness and it is too specific, as it considers only ST as a theory of mindreading and experientialism as a theory of the meaning of mental states.

In this paper, I want to extend Goldman's approach by focusing explicitly on ICs of preference strength and by considering the main theories of mindreading and of the meaning of mental states. I shall pursue a twofold goal. First, I shall individuate the conditions that ought to be satisfied in order to have reliable ICs of preference strength with respect to all the main accounts of how ordinary people make ICs of preference strength. Second, I shall examine some arguments attempting to show that these conditions can, at least in principle, be satisfied. By so doing, I believe it is possible to have a full assessment of the strategy under consideration. Ultimately, I shall claim that this strategy is unsuccessful. I shall offer an argument by elimination. No matter which account we adopt of how ordinary people make ICs of preference strength, all the arguments proposed in the literature fail to show that we can make reliable ICs of preference strength. Obviously, my analysis does not entail that no such argument exists. Nevertheless, the current state of research about ICs legitimates a moderate form of scepticism: given that no solution explored so far proves to be successful, it might as well be the case that ICs of preferences present an unsolvable problem.

I shall proceed as follows. In section 2, I shall illustrate the problem of ICs of preferences in more detail. Following Goldman, I shall specifically focus on its epistemological dimension, according to which the problem is whether or not we can have scientific knowledge of, or, at least, scientifically justified beliefs about, how different people's preferences compare in terms of strength. In section 3, I shall present, respectively, the main theories of mindreading and the main theories of the meaning of mental states present in contemporary philosophy of mind. In section 4, I shall consider the issue of scientific justification. Although scientific justification may require the satisfaction of several requirements (e.g. publicity, replicability, measurability), for simplicity here I shall focus only on one of them, namely, the requirement that the relevant mechanisms on which ordinary people's comparative practice is based be reliable for the purpose of making ICs. In section 6, I shall discuss five different arguments attempting to show that this condition can be satisfied. I shall argue that all these arguments fail. I shall summarise my findings in the conclusion.

2. THE PROBLEM

The problem of ICs of preferences presents several dimensions. However, two of them are especially important for the previously mentioned areas of scientific research. The first is the metaphysical dimension. The relevant question is whether or not there are facts about ICs of preference strength. The second is the epistemological dimension. The relevant question is whether or not we can have epistemic access to these (alleged) facts about ICs of preference strength. These dimensions are often confused in the literature. However, they differ in important respects and they should be distinguished. The metaphysical question is clearly prior to the epistemological question. If there is no fact of the matter about ICs of preference strength, then no question of epistemic access arises. Although the former is a contentious issue, in what follows, I shall simply presuppose that we can positively address the metaphysical question and focus instead on the epistemological question about ICs. Broadly speaking, this is the question of whether or not we can have knowledge of or, at least, justified, ICs of preference strength. More narrowly, it is the question of whether or not we

⁶ See A. Goldman, *Simulation and Interpersonal Utility*, «Ethics», 4, 1995, pp. 709-726.



can have ‘scientific’ knowledge of, or ‘scientifically’ justified beliefs about, how different people’s preferences compare in terms of strength. This issue is particularly important. Indeed, if we cannot have scientific knowledge of, or scientifically justified beliefs, about how different people’s preferences compare in terms of strength, welfare economics, social choice theory, normative and applied ethics remain in trouble even if there is a fact of the matter about ICs of preferences. It is thus important to see why ICs of preference strength raise specific epistemological worries.

As the problem of ICs of preferences often arises in the context of economics, it is useful to start our illustration from this area of analysis. In economics, preferences are defined as binary relations R , that is, relations between two items. The items included in the preference domain vary according to different decision theories.⁷ For the purpose at stake, it is not necessary to commit to any specific ontology. In what follows, I shall refer to the object of preference as the option that an individual faces. We can then say that an individual prefers an option x to another option y and express this in the following way: $x R y$. If preferences satisfy certain axioms, they can be represented by a numerical function. More specifically, if they satisfy the ordering axioms, i.e. completeness and transitivity, they can be represented by an ordinal utility function, unique up to a monotone increasing transformation. If preferences satisfy the expected utility axioms, i.e. the ordering axioms, the independence axiom and the Archimedean axiom, they can be represented by an interval utility function, unique up to a positive affine transformation. An interval function is a cardinal function. While an ordinal function is supposed to preserve simply the order of the individual’s preferences, a cardinal function is supposed to preserve additional information. More specifically, a cardinal function is supposed to capture the degree to which an individual prefers one option rather than another, or, in other words, the intensity of the individual’s preferences for alternative options. Typically, in order to fix a cardinal scale of measurement, it is (necessary and) sufficient to fix two points on the scale, namely, the zero and the unit. One feature of an interval scale of measurement is that both the zero and the unit are ‘arbitrary’. Thus, assigning the utility value 0 to one option does not mean that the individual prefers that option with zero intensity. It is simply an arbitrary choice, which fixes one of the relevant points on the cardinal scale of measurement.

Let us now consider an example. Suppose there are two individuals, i and j , and four options $x, y, w, z \in A$. Individual i ranks the options in the following way: $x R_i y R_i w R_i z$. On the other hand, individual j ranks the options in the following way: $w R_j z R_j x R_j y$. Suppose we represent their preferences on an interval scale. In particular, we assign the value 1 to the most preferred option and the value 0 to the worst option in their preference rankings. We can then assign a value that represents the intensity of their preferences for the other options, where these values are relative to the best and the worst options in each individual’s ranking. Suppose it is the case that $u_i(y) = u_j(x) = 0.6$. Suppose it is also the case that $u_i(x) - u_i(y) = u_j(w) - u_j(x) = 0.4$. In the first case, can we conclude that individual i prefers option y with the same strength with which individual j prefers option x ?⁸ In the second case, can we conclude that

⁷ More specifically, preferences range over either acts, or propositions, or prospects. For acts, see L. Savage, *The Foundations of Statistics*, Wiley, New York 1954. For propositions, see R. Jeffrey, *The Logic of Decision*, 2nd ed., University of Chicago Press, Chicago 1983 (1965, 1st ed.). For prospects, or lotteries, see J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton 1944.

⁸ This is the case of an interpersonal comparison of utility levels. ICs of utility levels are judgments of the form: $u_i(x) \geq u_j(y)$.



the difference in strength of individual i 's preference for option x over y is the same as the difference in strength of individual j 's preference for option w over z ?⁹

The answer to both questions is negative: the empirical evidence is not sufficient to conclude that identical utility values represent identical preference strengths. The reason is the following. As we have seen before, in order to fix an interval scale of measurement it is (necessary and) sufficient to fix an arbitrary zero and an arbitrary unit. In the case of preferences, we fix the scale of measurement by assigning the value 0 to the worst option and the value 1 to the best option. However, this is only sufficient to fix the scale of measurement for 'each' individual. It is not sufficient to fix a 'common' scale for both individuals. In order for the preference scale to be a common one, it must be the case that both individuals prefer their best option and their worst option with the same strength. The problem is that the evidence does not tell us anything at all about how different individuals' preferences for their best and worst options compare in terms of strength. As it is typically put, the evidence is consistent with the case where individual i prefers the best (the worst) option with intensity ten times greater than j .¹⁰

The previous example shows that ICs of preferences are underdetermined by the empirical evidence. At first sight, this poses a threat to the possibility of having scientifically justified ICs. At least, this seems to be the case if we adopt an evidentialist theory of epistemic justification.¹¹ The argument is straightforward. If all the possible empirical evidence is insufficient to determine ICs of preference strength, then, if the empirical evidence is what makes ICs of preference strength justified, it follows that ICs cannot be justified. On the other hand, if this is the only source of the problem, the epistemological problem of ICs might as well have a positive solution.¹² One reason is that evidentialism is not the only theory of epistemic justification. According to reliabilism, for instance, what makes a belief justified is not the empirical evidence, but the reliability of the processes by means of which the belief in question is formed.¹³ Thus, even if all the possible empirical evidence is insufficient to determine ICs of preference strength, these can nonetheless be justified, provided that they are acquired through reliable processes, that is, through processes that tend to produce true beliefs. The result is that, if we have independent grounds to prefer reliabilism to evidentialism as a theory

⁹ This is the case of an interpersonal comparison of utility differences. ICs of utility differences are judgments of the form: $u_i(x) - u_i(y) / u_j(w) - u_j(z) = \lambda$, for some $\lambda \in \mathfrak{R}$.

¹⁰ It is worth noticing that the problem is independent from which empirical evidence one considers. Economists typically take choice behaviour, under conditions of both certainty and uncertainty, to be the only admissible evidence for the ascription of individual preferences. However, as List has recently shown (See C. List, *Are Interpersonal Comparisons of Utility Indeterminate?*, «Erkenntnis», 58, 2003, pp. 229-260), it is not possible to compare different individuals' preference strengths in a meaningful way even if we extend the set of admissible evidence by including the individuals' latency of choice, i.e. the time delay between the presentation of the option and the actual choice-making, the probability of choice, i.e. the probability of choosing one option rather than another the probability of choice (See I. Waldner, *The Empirical Meaningfulness of Interpersonal Utility Comparisons*, «The Journal of Philosophy», 4, 1972, pp. 87-103), their verbal expressions (See J. Harsanyi, *Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility*, «The Journal of Political Economy», 63, 1955, pp. 309-321), their expressive reactions (See R. Weintraub, *Do Utility Comparisons Pose a Problem?*, «Philosophical Studies», 92, 1998, pp. 307-319), their facial expressions, body temperature and other proxies (See C. List, *op. cit.*).

¹¹ For a paradigmatic statement of this position, see R. Feldman and E. Conee, *Evidentialism*, «Philosophical Studies», 48, 1985, pp. 15-34.

¹² See A. Goldman, *Simulation and Interpersonal Utility*, *op. cit.*

¹³ For a paradigmatic statement of this position, see A. Goldman, *What is Justified Belief?*, in *Justification and Knowledge*, ed. by G. Pappas, Kluwer Academic Publisher, Reidel 1979, pp. 1-23.



of epistemic justification, we might be able to solve the epistemological problem of ICs, even if ICs are underdetermined by all the possible empirical evidence.

Let us go back to the goal of this paper. I said in the introduction that I am interested in assessing a specific strategy, which attempts to find a solution to the main theoretical problems concerning ICs of preference strength by investigating how ordinary people make ICs of preferences in everyday life. It is now clear that the problem about ICs of preference strength with which I am concerned is the problem of whether or not we can have scientific knowledge of, or scientifically justified beliefs about, how different people's preferences compare in terms of strength. In this paper, I shall adopt reliabilism as my theory of epistemic justification. According to reliabilism, in order for a belief to be justified, it is not necessary that we know that it is reliably acquired. It is only sufficient that, as a matter of fact, such a belief is reliably acquired. However, for other purposes, this may not be enough. For instance, if we want to give ICs of preference strength new scientific legitimacy, it may indeed be important to know whether or not ICs of preferences can be reliably made. The relevant question is thus whether or not there is scientific evidence that ICs of preference strength can be reliably made. As the strategy under consideration specifically focuses on how ordinary people make ICs of preferences in everyday life, this question can be further specified as the question of whether or not there is scientific evidence that ordinary people can reliably make ICs of preference strength in everyday life.¹⁴ This will be the object of my paper.

3. THEORIES OF MINDREADING

This section is divided in two parts. In the first, I shall briefly illustrate how the main theories of mindreading presented in the literature explain the ascription of mental states to a target. This involves considering what ordinary people mean when they use mental terms. In the second part, I shall attempt to clarify how these theories might explain the ascription to a target, and the interpersonal comparison, of preference strengths.

Let us start with the first part. Different disciplines formulate their explanatory accounts of people's mindreading capacity at different levels of description, i.e. the personal, the sub-personal and the physical level. Most of the philosophical literature is concerned with the sub-personal level of description. The common strategy is to conceive the mind as a system, where mental states and processes are characterized functionally. The goal is then to identify the underlying information-processing mechanisms that need to be postulated in order to explain our mindreading capacity. In the course of the years, two main approaches have emerged: Theory Theory (TT) and Simulation Theory (ST). TT characteristically accounts for the mindreading capacity by positing cognitive processes that exploit "an internally represented "knowledge structure" - typically a body of rules or principles or propositions - which serves to guide the execution of the capacity to be explained".¹⁵ In short, TT explains mental ascription

¹⁴ See A. Goldman, *Simulation and Interpersonal Utility*, *op. cit.*

¹⁵ See S. Stich and S. Nichols, *Folk Psychology: Simulation or Tacit Theory?*, «Mind and Language», 7, 1992, pp. 35-36. See also S. Stich and S. Nichols, *Second Thoughts on Simulation*, in *Folk Psychology: The Theory of Mind Debate*, ed. by M. Davies – T. Stone, Blackwell, Oxford 1995, pp. 87-108, S. Stich and S. Nichols, *Cognitive Penetrability, Rationality and Restricted Simulation*, «Mind and Language», 12, 1997, pp. 297-326, S. Nichols *et al.*, *Varieties of Off-Line Simulation*, in *Theories of Theories of Mind*, ed. by P. Carruthers – P. Smith, Cambridge University Press, Cambridge 1996, pp. 39-74, and S. Stich and S. Stich, *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, Oxford University Press, Oxford 2003.



by arguing that the folks possess a ‘Theory of Mind’ (ToM), to which they have a more or less conscious access.¹⁶

As far as the meaning of mental states is concerned, TT is generally associated with analytic functionalism. According to functionalism, the meaning of a mental state is given by the set of causal laws in which that mental state figures. Such causal laws specify how each mental state is related to environmental inputs, other mental states and behavioural outputs. It may be the case that the agent employing mental concepts is incapable of specifying all these constitutive causal relations. Indeed, this may require a sophisticated analysis. If we think of the defining causal relations as part of the ToM that the agent possesses, then we can say that such a theory operates tacitly, or, equivalently, that the theory is tacit. As far as mental ascription is concerned, then, TT assumes that the folks ascribe mental states to other people by observing external events (i.e. inputs and outputs) and inferring the relevant mental states by reference to the causal relations postulated by the ToM that they possess.

ST offers an alternative account of mental ascription. The basic idea is that mental ascription involves individuating another individual’s targeted mental states by imagining being subject to the same mental states to which she is subject. The ST approach to mindreading comes in different forms. In this paper, however, I shall primarily focus on the account advocated by Alvin Goldman.¹⁷ According to Goldman, first, the simulator asks herself what mental states she would have if she were subject to the initial mental states of the simulated agent. By so doing, she feeds her own information-processing mechanisms with pretend inputs, which supposedly correspond to the other person’s initial mental states. These

¹⁶ There are two variants of the TT approach to mindreading, namely, the scientific-theory theory (STT) and the modularity theory (MT). According to the former, the ToM that the folks use for mindreading is both learnt and stored in the mind in the same way as scientific theories are. In the course of their development, children proceed as little scientists, formulating hypotheses on the basis of the information available and revising them in the light of new data. According to the latter, the ToM is neither learnt nor stored in the same way as scientific theories are, but it is rather included in one or more innate modules. For the current purpose, however, we can ignore the distinction between the two approaches. See H. Wellman, *The Child’s Theory of Mind*, MIT Press, Cambridge, Mass. 1990, J. Perner, *Understanding the Representational Mind*, MIT Press, Cambridge, Mass. 1991, J. Gopnik and H. Wellman, *Why the child’s theory of mind really is a theory of mind*, «Mind and Language», 7, 1992, pp. 145-171, J. Gopnik and H. Wellman, *The theory theory*, in *Mapping the Mind: Domain Specificity in Cognition and Culture*, ed. by L. Hirschfeld – S. Gelman, Cambridge University Press, New York 1994, J. Gopnik and A. N. Meltzoff, *Words, Thoughts and Theories*, MIT Press, Cambridge, Mass. 1997, for a defence of the STT approach. See A. Leslie, *Pretence and representation: The origins of “theory of mind”*, «Psychological Review», 94, 1987, pp. 412-426, A. Leslie, *Some implications of pretense for mechanisms underlying the child’s theory of mind*, in *Developing Theories of Minds*, ed. by J. Astington – P. Harris – D. Olson, Cambridge University Press, Cambridge 1988, pp. 19-46, A. Leslie, *Pretending and Believing: Issues in the theory of ToMM*, «Cognition», 50, 1994, pp. 211-238, A. Leslie, *How to acquire a representational theory of mind*, in *Metarepresentation: A Multidisciplinary Perspective*, ed. by D. Sperber, Oxford University Press, New York 2000, pp. 197-223, A. Leslie and T. German, *Knowledge and ability in “theory of mind”: One-eyed overview of a debate*, *Mental Simulation*, in ed. by M. Davies – T. Stone, Blackwell, Oxford 1995, pp. 123-150, and S. Baron-Cohen, *Mindblindness: An Essay on Autism and Theory of Mind*, MIT Press, Cambridge, Mass. 1995, for a defence of the MT approach.

¹⁷ See, in particular, A. Goldman, *Interpretation Psychologized*, «Mind and Language», 4, 1989, pp. 161-185, A. Goldman, *In defense of the simulation theory*, «Mind and Language», 7, 1992, pp. 104-119, A. Goldman, *The mentalizing folk*, in *Metarepresentations*, ed. by D. Sperber, Oxford University Press, Oxford 2000, A. Goldman, *Simulation theory and mental concepts*, in *Simulation and Knowledge of Action*, ed. by J. Dokic – J. Proust, John Benjamins, Amsterdam 2002, pp. 1-20, A. Goldman, *Simulating Minds*, Oxford University Press, Oxford 2006.



mechanisms run ‘off-line’ and produce pretend outputs. The simulator then introspects these mental states and recognises them as belonging to a certain type. Finally, she ascribes them – by analogy – to the simulated agent.¹⁸

One crucial feature of Goldman’s account is its emphasis on introspection. Indeed, third-personal mental ascription presupposes that the simulator is capable of introspecting her own mental states in order to ascribe mental states to another individual. If this is the case, the ST explanation of mindreading requires an account of ordinary people’s introspective capacity. One suggestion is that the simulator recognises her own mental states with respect to the function that they occupy in her mind-system. Goldman gives two reasons to reject this option. The first is that functionalism characterises mental states as dispositions, which may involve relationships with events to which the person has not yet epistemic access (e.g. future events) or will never have epistemic access (e.g. in the case of subjunctive relationships). The second is that, in order for an individual to recognise her own occurrent mental states, functionalism requires that she be able to identify the indefinitely large number of other attitudes to which her occurrent mental states are related as a matter of definition. In turn, this seems to burden self-ascription with excessive computational requirements.¹⁹

In the light of these problems, Goldman suggests that a more plausible account explains self-ascription in terms of the capacity of internally detecting mental properties that are both categorical and non relational (or at least non-massively relational). There are two candidate types of properties satisfying these criteria: phenomenological and non-phenomenological properties. In his “Simulation and Interpersonal Utility”, Goldman opts for the former candidate. Accordingly, the simulator recognises her mental states on the basis of their phenomenology, that is, on the basis of ‘what it is like’ to have them in specific circumstances. Effectively, this means embracing an experientialist view of the meaning of mental states, that is “the traditional view that mental language gets its meaning, primarily and in the first instance, from episodes of conscious experience of which the agent is more or less directly aware”.²⁰ It follows that the corresponding ST account of mental ascription is associated with a view of mental states as phenomenologically real states, to which the agent has introspective – privileged, although not infallible – access.²¹

We can now examine more closely how the two main mindreading approaches explain the folks capacity to ascribe preferences to other individuals and to compare them in terms of strength. Let us consider TT first. As seen above, TT is typically associated with a functionalist understanding of mental states. Preferences can be defined in functionalist terms as mental states that are causally related to certain inputs, and that, in combination with other mental states, produce certain behavioural outputs. According to some authors, decision theory is the

¹⁸ See A. Goldman, *Interpretation Psychologized*, *op. cit.*

¹⁹ See A. Goldman, *The Psychology of Folk Psychology*, «Behavioral and Brain Sciences», 16, 1993, pp. 15-28, and A. Goldman, *Simulation theory and mental concepts*, *op. cit.*

²⁰ See A. Goldman, *Simulation and Interpersonal Utility*, *op. cit.*, p. 712.

²¹ See A. Goldman, *The Psychology of Folk Psychology*, *op. cit.* It is worth noticing that Goldman has recently changed his mind about the meaning of mental states. At present, he defends the view that mental concepts pick out categorical properties of mental states that are non-phenomenal properties. See A. Goldman, *Simulating Minds*, *op. cit.* In what follows, I shall still focus on the version of ST associated with experientialism for two reasons. The first is that this is the most developed version of ST. The second is that Goldman’s most recent account of the meaning of mental states is not yet elaborated in sufficient details to provide a basis for an accurate discussion of the problem of ICs of preference strength.



research programme that attempts to specify some of these relevant relations.²² In particular, decision theory conceives preferences as mental states that lead to choices, in combination with beliefs and desires.²³ Thus, if we define preferences in functionalist terms, the property of preferential strength can be conceived as the causally efficacious property, which leads an individual to behave in a certain way, when subject to specific circumstances and in the presence of other beliefs and desires, in accordance with the causal laws defining the notion of preference.

Suppose now that an observer, e.g. a judge, wants to compare another individual's preferences with her own in terms of strength. The first step concerns third-person mental ascription. The judge observes the relevant external events (i.e. instances of the input-types and output-types that are included in the definition of preference) and infers both the content and the strength of the other individual's relevant preferences, by reference to the causal relations postulated by the ToM that she possesses. The second step concerns first-person mental ascription. Orthodox TT suggests that first-person mental ascription entirely parallels third-person mental ascription. This means that self-ascription is based on inferences mediated by the ToM that the judge possesses. Less orthodox TT approaches relax this position by conjecturing that first-person mental ascription involves the use of recognitional devices or mechanisms – which either make the use of the ToM invisible, but not completely irrelevant, or confine it to certain specific purposes – and ends up with the self-ascription of both a specific content and a specific strength to one's own preferences. The last stage concerns the interpersonal comparison of preferences. The judge has formed a belief about the intensity of the other individual's preferences and a belief about the intensity of her own preferences. Straightforwardly, she can now combine those beliefs to make an interpersonal comparison of preference strengths.

Let us now consider ST. As seen above, Goldman's ST version is associated with an experientialist understanding of mental states.²⁴ Preferences can be defined in experientialist terms as mental states that give rise to certain experiences in a subject. It may be the case that there is no unique phenomenal experience that different individuals have in common when they are in a preference-state. However, it is enough that there is a family of experiences that are sufficiently similar to constitute a preference-type. According to an experientialist understanding, then, preference strength is a felt property, a qualitative experience of the individual that has preferences. The subject has introspective access and can discriminate the strengths of his preferences. As such, preference strength is a real psychic magnitude, whose

²² See D. Lewis, *Philosophical Papers. Vol. 1*, Oxford University Press, Oxford 1986, P. Pettit, *Decision Theory and Folk Psychology*, in *Foundations of Decision Theory*, ed. by M. Bacharach – S. Hurley, Basil Blackwell, Oxford 1991, pp. 147-175, and P. Pettit, *Preference, Deliberation and Satisfaction*, «Royal Institute of Philosophy Supplement», 81, 2006, pp. 131-154.

²³ Roughly speaking, there are three possible ways to conceive the relationship between desires and preferences. First, one can be eliminativist about preferences and claim that the notion of preferences is syncategorematic. It is simply a way to conveniently describe an individual's desires and their relations. However, there are no real mental states corresponding to preferences. Second, one can be reductivist and claim that preferences are real mental states but mental states that reduce to desires in one sense or another, e.g. they constitute a specific, e.g. relational, class of desires. Finally, one can maintain that preferences are derivative on desires, in the sense that they are related to, and determined by, them; but they do not reduce to desires, except in the loose sense that they are both pro-attitudes of some sort. I think that the functionalist position fits more comfortably with the latter position, which I shall thereby adopt in what follows.

²⁴ See particularly the account offered by A. Goldman, *Simulation and Interpersonal Utility*, *op. cit.*



meaning arises “from points or intervals on the experiential scale”²⁵ that the term denotes, and which the subject experiences and can introspectively discriminate.

Once again, suppose that one individual, e.g. the simulator, wants to compare another individual’s preferences with her own in terms of strength. The first step concerns third-person mental ascription. The simulator asks herself which content and intensity her preferences would have if she had the simulated agent’s initial mental states. This involves recreating in imagination the same qualitative experiences of the individual whose preference she wants to compare. Then, the simulator discriminates the intensity of these experiences through introspection and classifies them as experiences of preferences with a specific intensity. Lastly, she ascribes such preference strengths to the other agent by analogy. The second step concerns first-person mental ascription. Self-ascription proceeds by direct introspection. The simulator detects her own preferences, discriminates their intensities and ascribes them to herself. Finally, the last stage concerns the interpersonal comparison of preferences. Once again, the simulator has formed a belief about the intensity of the other individual’s preferences and a belief about the intensity of her own preferences. She can then combine those beliefs to make an interpersonal comparison of preference strengths.

It is worth noticing that both TT and ST set only minimal conditions for the ‘possibility’ of forming beliefs about how different people’s preferences compare in terms of strength. In both cases, the explanation of how ordinary people make ICs of preference strength is consistent with the possibility that these beliefs are systematically mistaken. Two questions arise. First, what conditions should be satisfied in order for ordinary people to make reliable ICs of preference strength, within a TT and a ST account of their mindreading capacity? Second, can these conditions, at least in principle, be satisfied?

4. CONDITIONS FOR RELIABLE ICS OF PREFERENCE STRENGTH

In this section I want to consider the first of the previous questions. Let us start with TT first. If ordinary people make ICs of preference strength as the TT approach suggests, there are two requirements that must be satisfied for such ICs to be reliably made. First, the judge’s inferences about preference strengths must be based on the correct inputs. Second, they must be based on the correct theory about the compared individuals’ mind.

The first requirement is straightforward. If the evidence that the judge uses is not correct, then she is likely to reach wrong conclusions about the observed agents’ preference strengths. The second requirement holds that the judge’s inferences must be based on the correct theory about the compared agents’ mind. This means that the causal laws that form the ToM that the judge possesses must correctly represent, or at least very closely approximate, the way in which the relevant information-processing mechanisms of the targeted agents work. For simplicity, let us refer to this requirement as the condition of ToM-to-mind similarity.

Let us now move to ST. Summarising Goldman’s own position and a large literature on ST, we can distinguish three requirements that must be satisfied for ICs to be reliably made. First, simulation must be based on the correct inputs. Second, the simulator and the simulated agents must be similar at the level of the relevant information-processing mechanisms. Third, the simulator’s relevant information-processing mechanisms must operate in the same way in imagination as in reality.

The rationale underlying the first requirement is identical to the TT case. If the simulator feeds his information-processing mechanisms with incorrect inputs, then she is likely to reach

²⁵ See A. Goldman, *Simulation and Interpersonal Utility*, *op. cit.*, p. 713.



wrong conclusions about the simulated agents' preference strengths. One problem arises. In Goldman's account, the inputs are pretend mental states corresponding to the agent's actual mental states. Thus, in order for simulation to be reliable, it must be the case that the simulator can correctly individuate the intensity of the agent's actual mental states. However, this task presents the same difficulties associated with the comparison of different individuals' preference strengths. Thus, the assumption that the simulator can feed the 'correct' inputs into his 'off-line' system simply begs the question.

In the light of this problem, it seems better to take the simulated agent's environmental circumstances, rather than pretend mental states, as inputs of simulation. The simulator does not begin by asking herself what preferences she would have if she were to have another individual's initial mental states. Instead, she begins by asking herself what preferences she would have if she were in the other individual's initial circumstances. ST must be thus complemented with causal knowledge about the relations between environment and mental states such as beliefs and desires. Moreover, it must be complemented with knowledge about the history of the simulated agent, which should be used to identify which environmental circumstances constitute relevant inputs in specific situations, amongst the infinite ones that the mere observation of the simulated agent's situation allows one to consider. Clearly, this moves ST towards a more hybrid formulation.

Consider now the second requirement, according to which the simulator and the simulated agent must be similar at the level of the relevant information-processing mechanisms. Let us refer to it as the assumption of interpersonal psychological similarity. The rationale of this requirement is intuitive. Even if the simulator feeds her information-processing mechanisms with the correct inputs, she will reach the wrong conclusions about the agent's mental states unless they are psychologically similar in the respects that matter for forming preference strengths.

Finally, consider the third requirement. Even if the assumption of interpersonal psychological similarity is satisfied, so that the simulator forms preferences that are similar to those of the target individual on the basis of 'actual' causal circumstances, it might still be the case that she forms preferences that are radically different from those of the other individual on the basis of 'imagined' causal circumstances. After all, simulation works 'off-line', whereas interpersonal psychological similarity is a thesis about 'on-line' information-processing mechanisms. In order for simulation to be reliable for ICs, the previous requirement must be complemented with the requirement that the off-line working of the individual's mind-system approximates its online working.

5. FIVE ARGUMENTS

In this section I want to examine whether or not the conditions for having reliable ICs of preference strength can be met. There is now considerable evidence that the third requirement for the reliability of simulation can be satisfied.²⁶ Thus, here, I shall mainly focus on the first and the second requirements for the reliability of both TT and ST. Following and extending Goldman's analysis, I shall consider five arguments, which I shall call, respectively, the argument from mindreading predictive success, the argument from neuroscience, the argument from evolution, the argument from scientific practice and the argument from nativism.

²⁶ See, for instance, G. Currie and I. Ravenscroft, *Mental Simulation and Motor Imagery*, «Philosophy of Science», 64, 1997, pp. 161-180.



The argument from mindreading predictive success claims that the fact that mindreading is reliable for predictive purposes provides *prima facie* evidence that mindreading is reliable also for the purpose of making IC judgments. The reason is that IC judgments are based on the same mental ascriptions that lead us to reliable behavioural predictions. The objection against this argument is that, even if we grant that mindreading is reliable for predicting an agent's behaviour, we cannot conclude that mindreading is also reliable for making ICs of preference strength. Let us see why by considering TT and ST in more detail.

Consider TT first. The judge ascribes preference strengths to herself and to the target, which best predict their respective behaviour on the basis of the available evidence. The problem is that it is possible for the judge to make both correct behavioural predictions and incorrect ICs of preference strength. For instance, the following is one possible reason. As preferences are characterised in functionalist terms, their intensity is relative to the intensity of other mental states, e.g. other preferences. However, such mental states may never become manifest in overt behaviour. It follows that two individuals' preference strengths may appear identical, from the point of view of all the possible empirical evidence, and yet be different, insofar as they are relative to the intensity of mental states that never generate behavioural outputs. Clearly, the judge can make successful behavioural predictions even if not all of the agent's mental states become manifest in overt behaviour. On the other hand, her mindreading activity may be unreliable for making ICs, because there is no way to show that the inputs that are relevant for ICs are really correct.

Consider ST now. Once again, reliable behavioural predictions are consistent with different and incompatible IC judgments. This means that although it may be true that "empirically observed success at empathy-based predictions of behaviour does go some distance toward supporting psychological isomorphism",²⁷ it is not true that predictive success goes far enough in showing that such a psychological isomorphism is high enough to lead to correct ICs of preference strength. One problem is that the simulator and the simulated agent may differ with respect to one of the relevant information-processing mechanisms, despite the fact that this difference never becomes manifest at the predictive level. For instance, it may be possible that the simulated agent responds to the environmental inputs by forming desires with intensity ten times greater than the simulator's. If the evidence about the two individuals is perfectly identical, the simulator can make successful predictions of the other agent's behaviour and yet incorrect ICs of preference strength.

To summarise, success at predicting an agent's behaviour requires both a less fine-grained individuation of the relevant inputs and a looser degree of similarity than those required for having reliable ICs of preference strength. At best, predictive success shows that mindreading is reliable for predictive purposes. However, it does not offer a reason to think that mindreading is reliable also for making ICs of preference strength.

The argument from neuroscience claims that, if it is possible to establish well-defined correlations between the intensity of the judge's neural activation during preference ascription and the intensity of the agent's neural activation during preference formation, then it is possible to claim that the mechanisms underlying mindreading are isomorphic to the mechanisms underlying preference formation. This argument specifically fits the ST mindreading approach. Indeed, as simulation employs 'off-line' the same mechanisms that are activated during 'on-line' preference formation, the hypothesised neural correlation should figure amongst the predictions of the theory. If a correlation of that sort can be robustly

²⁷ See A. Goldman, *Simulation and Interpersonal Utility*, *op. cit.*, p. 724.



established, it may thus be possible to vindicate the assumption of interpersonal psychological similarity.

This argument faces some problems. To begin with, although there is evidence that some mental states, e.g. disgust, are located in specific brain regions and, thereby, that different individuals undergoing those states share common neural properties, the same is not true for other mental states, like preferences. Perhaps, this is simply a problem of limited empirical evidence. It might as well be the case that one day scientific research will discover the neural correlates of preferences. If so, the argument from neuroscience may enjoy a good fate. Even if we grant this possibility, however, I believe that the prospects of success are dim. The existence of a common neural region dedicated to preference formation does not imply, *per se*, that an identical neural activation corresponds to the formation of preferences with identical strength across individuals.

Consider the experientialist understanding of preferences associated with ST. If we grant the possibility that the qualitative character of experiences is not fully accounted by their neurophysiological character, it is clear that identical neuronal activation across individuals may correspond to preference experiences that are significantly different at the level of strength. This is the same as admitting that interpersonal isomorphy at the physical level does not necessarily imply interpersonal isomorphy at the subjective level.

Things do not change if we adopt a functionalist understanding of preferences. According to it, preference strengths are individuated not only with respect to external inputs and outputs, but also with respect to other mental states. Crucially, these states can be both occurrent and non-occurrent. The problem is that neural activation registers only occurrent mental states. In order to conclude that different individuals' preference strengths are the same when their neurons fire with the same intensity, we need to assume that they are identical with respect to all those non-occurrent mental states which might impact on their occurrent preference strengths. However, we have no epistemic reason to accept this *ceteris paribus* assumption. Once again, the result is that interpersonal isomorphy at the physical level does not imply interpersonal isomorphy at the functional level.

If the previous points are correct, the hypothesised correlation between the judge's mindreading mechanisms and the agent's preference formation mechanisms does not prove that ICs can be reliably made. The reason is the following. Although the activation of the neural region dedicated to mindreading may co-vary with the activation of the neural region dedicated to preference formation, it may still be the case that the judge does not get the other individual's preference strengths right. In other words, interpersonal neural correlation may be consistent with systematic errors about preference strength attribution. Therefore, the argument from neuroscience does not support the hypothesis that ICs of preference strength can be reliably made.

The argument from evolution claims that evolutionary pressure might have favoured the development of a close isomorphism between the observer's ToM and the target's information-processing mechanisms, in the case of TT, or between the simulator's and the simulated agent's information-processing mechanisms, in the case of ST. The reason is that this would have maximised the expected fitness of the members of a relevant group by endowing them with competitively advantageous features for the typical environment encountered by the group. Roughly speaking, individual fitness is assessed with respect to the (probabilistic) propensity of spreading one's own genes into the next generation. Clearly, this propensity may be enhanced by the individual's capacity to correctly predict, explain or interpret other group members' behaviour. If this is the case, evolutionary pressure might have favoured the development of a strong convergence in the working of different individuals' mindreading mechanisms and, thereby, a high level of intersubjective agreement about the outcomes of



mindreading. The problem is that, as I shall illustrate below, intersubjective agreement presupposes a less demanding degree of similarity than the one required for having reliable IUC judgments. Given that a higher degree of similarity would be unnecessarily costly, it follows that the argument from evolution does not support the claim that ICs of preference strength can be reliably made.

Let us consider TT first. Suppose there are two individuals trying to mindread each other. Suppose that the ToM that each of them uses closely represents the working of the other individual's mind. Finally, suppose that the empirical evidence about their preferences is identical. On the basis of the ToM in their possession, they will both conclude that they have the same preference strengths. However, it may as well be the case that the first individual's occurrent preference strengths are relative to a different set of non-occurrent mental states, which never become manifest. The consequence is that intersubjective agreement is consistent with the case where ICs of preferences are unreliably made. Intersubjective agreement merely requires that the individuals make similar assumptions about a suitable set of non-occurrent mental states, i.e. those that have a chance of becoming manifest in overt behaviour.

Let us now consider ST. Suppose that there are two individuals simulating each other's mental life. Suppose that both individuals are completely identical at the level of the relevant information-processing mechanisms, except for the fact that the first individual forms desires with intensity ten times greater than the second individual's, when responding to the same environmental stimuli. If the relevant evidence is the same, both individuals will conclude that they have identical preference strengths. Indeed, both individuals ascribe preferences to the other on the basis of their own cognitive machinery, under the assumption that the other individual is similar to them in the relevant respects. However, by so doing, they make incorrect ICs of preference strength, because, by stipulation, the intensity of the first individual's desires is ten times greater than the intensity of the second individual's desires. Intersubjective agreement requires a looser degree of interpersonal psychological similarity than the one required for making reliable ICs.

The argument from scientific practice starts from the observations that, on the one hand, most scientific theories are underdetermined by the empirical evidence, like ICs of preference strength, and yet that, on the other hand, there are often reasons to prefer one theory rather than another on grounds of simplicity, parsimony and non-arbitrariness. These are pragmatic virtues that characterise scientific practice and that help break the underdetermination of scientific theories by the empirical evidence. Applied to ICs of preference strength, the argument works in the following way. Suppose that two different judges, in the case of TT, or simulators, in the case of ST, reach intersubjective agreement about other individuals' preference strengths. It is true that preference strengths may be relative to other mental states that never become manifest or that there may be hidden interpersonal differences concerning the relevant information-processing mechanisms, but if all the empirical evidence is otherwise the same, the best explanation of the mindreaders' intersubjective agreement is that the other individuals' preference strengths are really the same. In both cases, the recommended conclusion looks like the simplest, the most parsimonious and the least arbitrary hypothesis.

The objection against this line of thought is that it contrasts with the very explanatory practice in the mindreading literature. In the case of TT, the best explanation of intersubjective agreement in mindreading is that different judges 'take' their theories about other people's mind to be correct, on the one hand, and the *ceteris paribus* assumption about non-occurrent mental states to be satisfied, on the other hand. This does not mean that the ToM in their



possession or the inputs that they consider are really correct. Likewise, in the case of ST, the best explanation of intersubjective agreement is that different simulators ‘take’ other individuals to be just like them at the level of the relevant information-processing mechanisms. Once again, this does not mean that simulators and simulated agents are really alike. In other words, the mindreading literature explains intersubjective agreement by holding that the judges or the simulators merely ‘take’ the reliability requirements to be satisfied, without holding that they really are. If this the case, the argument from scientific practice does not show that ICs of preference strength can be reliably made.

The argument from nativism is specifically advocated by Goldman and pursues an analogy with Chomsky’s nativist approach in linguistics.²⁸ Chomsky’s analysis starts from the observation that children belonging to the same community end up acquiring the same grammar. This fact is particularly striking because grammar acquisition is radically underdetermined by the empirical evidence. According to Chomsky, it is not plausible to assume that children use purely pragmatic criteria, such as simplicity and parsimony, in order to learn a common grammar amongst the infinitely many possible ones that are consistent with the available empirical evidence. Instead, Chomsky suggests that children possess an innate and universal body of knowledge, which guides them in the process of language learning. Such a body of knowledge is not only important during the development process. Indeed, it is the very body of knowledge on which the grammaticality judgments of adult competent speakers are based.

Goldman invites us to conceive the problem of ICs in analogy with linguistics. The starting point is the observation that different observers reach frequent intersubjective agreement about ICs of preference strength. This fact is particularly striking because, as we have seen, ICs are radically underdetermined by the empirical evidence. As the analogy with linguistics suggests, it is not plausible to assume that different observers form the same beliefs, amongst the infinite ones licensed by the empirical evidence, on the basis of purely pragmatic considerations.²⁹ Rather, it is more plausible to hold that they form the same beliefs on the basis of the possession an innate and highly representative ToM, in the case of TT, and innate and highly similar information-processing mechanisms, in the case of ST. According to Goldman, if the nativist hypothesis gives linguistics “epistemic respectability”, so does it with ICs of preference strength.³⁰

The crucial concept is that of ‘innateness’. The question of what innateness is has generated a particularly intense philosophical debate in the past few years.³¹ Although there is an evident lack of agreement in the literature, the most recent positions suggest taking ‘nativism’ as equivalent to ‘psychological primitivism’.³² Roughly speaking, innate cognitive capacities are psychological primitives. In turn, psychological primitives are entities or processes that, on the one hand, are mentioned in the best psychological explanations of human behaviour; and, on the other hand, whose acquisition cannot be explained by any psychological theories, but only by a theory at a lower level. The important issue is to see

²⁸ See N. Chomsky, *Rules and representations*, Basil Blackwell, Oxford 1980.

²⁹ Cfr. J. Harsanyi, *Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility*, op. cit, and J. Harsanyi, *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*, Cambridge University Press, Cambridge 1977.

³⁰ See A. Goldman, [1995], pp. 725-726.

³¹ See F. Cowie, *What’s Within? Nativism Reconsidered*, Oxford University Press, Oxford 1999, P. Griffiths, *What is innateness?*, «Monist», 85, 2002, pp. 70 – 85, R. Samuels, *Nativism in cognitive science*, «Mind & Language», 17, 2002, pp. 233 – 265, M. A. Khalidi, *Innate Cognitive Capacities*, «Mind and Language», 22, 2007, pp. 92-115.

³² See specially F. Cowie, op. cit, and R. Samuels, op. cit.



whether or not there is scientific evidence supporting the nativist hypothesis in the case of ICs of preference strength.

The first objection is that, even if such evidence exists, it does not show that ICs of preference strength can be reliably made. After all, nativism is a hypothesis about the acquisition of the mindreading capacity rather than a hypothesis about its reliability. In order to turn the argument from nativism into an argument of the right kind, we should reformulate it as claiming that the reliability of the relevant information-processing mechanisms constitutes a ‘primitive’ in a theory of how ordinary people make ICs of preference strength, in the sense that the best explanation of their comparative capacity and of their frequent intersubjective agreement requires such property. In the light of the previous discussion, however, we can easily reject even this alternative version of Goldman’s argument. As we have seen above, the best explanation of people’s comparative practice and intersubjective agreement merely requires the assumption that ordinary people ‘take’ the reliability requirements to be satisfied, not they really are. Therefore, even the argument from nativism fails.

6. CONCLUSION

Solving the problem of ICs of preference strength is of vital importance for welfare economics, social choice theory and ethics. It is therefore not surprising that several attempts have been made in the course of the years. In this paper, I examined one strategy in particular, which tries to solve the problem of ICs by looking more closely at how ordinary people make ICs of preference strength in everyday life. The question that I specifically considered is whether or not there is scientific evidence that ordinary people make ICs of preference strength in a reliable way. I discussed five arguments, which offer various reasons to think that the answer to this question is affirmative. In this paper, I argued that they all fail. This does not mean that no positive argument exists, which can vindicate the strategy under scrutiny or, more generally, solve the problem of ICs. However, given that all the solutions explored so far turn out to be unsuccessful, this paper provides *prima facie* reason to think that scepticism about ICs may be here to stay.

BIBLIOGRAPHY

- Arrow K. (1995) *Social Choice and Individual Values*, 2nd ed., Wiley, New York 1963 (1951, 1st ed.).
- Baron-Cohen, S. *Mindblindness: An Essay on Autism and Theory of Mind*, MIT Press, Cambridge, Mass., for a defence of the MT approach.
- Chomsky N. (2002) *Rules and representations*, Basil Blackwell, Oxford 1980.
- Cowie, F. *What’s Within? Nativism Reconsidered*, Oxford University Press, Oxford 1999, P. Griffiths, *What is innateness?*, *Monist*, 85, pp. 70 – 85.
- Currie G. and Ravenscroft I. (1997) Mental Simulation and Motor Imagery, *Philosophy of Science*, 64, pp. 161-180.
- Davidson D. (1986) Judging interpersonal interests, in *Foundations of social choice theory*, ed. by J. Elster – A. Hylland, Cambridge University Press, Cambridge, pp. 195-211.



- Davidson D. (2004) *Problems of Rationality*, Oxford University Press, Oxford.
- Feldman R. and Conee E. (1985) *Evidentialism*, *Philosophical Studies*, 48, 1985, pp. 15-34.
- Goldman A. (1979) *What is Justified Belief?*, in *Justification and Knowledge*, ed. by G. Pappas, Kluwer Academic Publisher, Reidel, pp. 1-23.
- Goldman A. (1989) Interpretation Psychologized, *Mind and Language*, 4, pp. 161-185.
- Goldman A. (1992) In defense of the simulation theory, *Mind and Language*, 7, pp. 104-119.
- Goldman A. (1993) The Psychology of Folk Psychology, *Behavioral and Brain Sciences*, 16, pp. 15-28.
- Goldman A. (1995) *Simulation and Interpersonal Utility*, *Ethics*, 4, 1995, pp. 709-726.
- Goldman A. (2000) *The mentalizing folk*, in *Metarepresentations*, ed. by D. Sperber, Oxford University Press, Oxford.
- Goldman A. (2002) Simulation theory and mental concepts, in *Simulation and Knowledge of Action*, ed. by J. Dokic – J. Proust, John Benjamins, Amsterdam, pp. 1-20.
- Goldman A. (2006) *Simulating Minds*, Oxford University Press, Oxford.
- Gopnik, J. and Wellman H. (1992) *Why the child's theory of mind really is a theory of mind*, *Mind and Language*, 7, pp. 145-171.
- Gopnik J. and Wellman H. (1994) *The theory theory*, in *Mapping the Mind: Domain Specificity in Cognition and Culture*, ed. by L. Hirschfeld – S. Gelman, Cambridge University Press, New York.
- Gopnik J. and Meltzoff A. N. (1997) *Words, Thoughts and Theories*, MIT Press, Cambridge, Mass.
- Harsanyi J. (1955) Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility, *The Journal of Political Economy*, 63, pp. 309-321.
- Harsanyi J. (1977) *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*, Cambridge University Press, Cambridge.
- Jeffrey R. (1983) *The Logic of Decision*, 2nd ed., University of Chicago Press, Chicago (1965, 1st ed.).
- Jevons S. (1911) *Theory of Political Economy*, 4th edition, Macmillan, London (1871, 1st ed.).
- Khalidi, M. A. (2007) Innate Cognitive Capacities, *Mind and Language*, 22, 2007, pp. 92-115.
- Leslie A. (1987) Pretence and representation: The origins of "theory of mind", *Psychological Review*, 94, 1987, pp. 412-426.
- Leslie A. (1988) *Some implications of pretense for mechanisms underlying the child's theory of mind*, in *Developing Theories of Minds*, ed. by J. Astington – P. Harris – D. Olson, Cambridge University Press, Cambridge, pp. 19-46.
- Leslie A. (1994) Pretending and Believing: Issues in the theory of ToMM, *Cognition*, 50, , pp. 211-238.
- Leslie A. (2000) How to acquire a representational theory of mind, in *Metarepresentation: A Multidisciplinary Perspective*, ed. by D. Sperber, Oxford University Press, New York, pp. 197-223.



- Leslie A. and German T. (1995) *Knowledge and ability in "theory of mind": One-eyed overview of a debate, Mental Simulation*, in ed. by M. Davies – T. Stone, Blackwell, Oxford, pp. 123-150.
- Lewis D. (1986) *Philosophical Papers. Vol. 1*, Oxford University Press, Oxford.
- List L. (2003) Are Interpersonal Comparisons of Utility Indeterminate?, *Erkenntnis*, 58, pp. 229-260.
- Little I.D.M. (1957) *A Critique of Welfare Economics*, 2nd ed., Clarendon Press, Oxford (1950, 1st ed.).
- Nichols S. et al. (1996) *Varieties of Off-Line Simulation*, in *Theories of Theories of Mind*, ed. by P. Carruthers – P. Smith, Cambridge University Press, Cambridge, pp. 39-74.
- Nichols, S. and Stich, S. (2003) *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, Oxford University Press, Oxford.
- Perner J. (1991) *Understanding the Representational Mind*, MIT Press, Cambridge, Mass.
- Pettit P. (1991) *Decision Theory and Folk Psychology*, in *Foundations of Decision Theory*, ed. by M. Bacharach – S. Hurley, Basil Blackwell, Oxford, pp. 147-175.
- Pettit P. (2006) Preference, Deliberation and Satisfaction, *Royal Institute of Philosophy Supplement*, 81, 2006, pp. 131-154.
- Robbins L. (1932) *An Essay on the Nature and Significance of Economic Science*, Macmillan, London.
- Samuels R. (2002) Nativism in cognitive science, *Mind & Language*, 17, pp. 233 – 265.
- Savage L.(1954) *The Foundations of Statistics*, Wiley, New York.
- Stich S. and Nichols S. (1992) Folk Psychology: Simulation or Tacit Theory?, *Mind and Language*, 7, pp. 35-36.
- Stich S. and Nichols S. (1995) *Second Thoughts on Simulation*, in *Folk Psychology: The Theory of Mind Debate*, ed. by M. Davies – T. Stone, Blackwell, Oxford, pp. 87-108.
- Stich S. and Nichols S. (1997) Cognitive Penetrability, Rationality and Restricted Simulation, *Mind and Language*, 12, 1997, pp. 297-326.
- von Neumann J. and Morgenstern O. (1944) *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.
- Waldner I. (1972) *The Empirical Meaningfulness of Interpersonal Utility Comparisons*, *The Journal of Philosophy*, 4, pp. 87-103.
- Weintraub R. (1998) Do Utility Comparisons Pose a Problem?, *Philosophical Studies*, 92, pp. 307-319.
- Wellman H. (1990) *The Child's Theory of Mind*, MIT Press, Cambridge, Mass.