JOURNAL OF APPLIED ETHICS AND PHILOSOPHY

Center for Applied Ethics and Philosophy
Hokkaido University

vol. 3

September 2011

CONTENTS

Editorial Note

The Journal of Applied Ethics and Philosophy is an interdisciplinary periodical covering diverse areas of applied ethics. It is the official journal of the Center for Applied Ethics and Philosophy (CAEP), Hokkaido University. The aim of the *Journal of Applied Ethics and Philosophy* is to contribute to a better understanding of ethical issues by promoting research into various areas of applied ethics and philosophy, and by providing researchers, scholars and students with a forum for dialogue and discussion on ethical issues raised in contemporary society.

The journal welcomes papers from scholars and disciplines traditionally and newly associated with the study of applied ethics and philosophy, as well as papers from those in related disciplines or fields of inquiry.

Earlier versions of the papers by Søren Holm and Bernard Baertschi published in this present volume of the *Journal of Applied Ethics and Philosophy* were delivered at the Fifth International Conference on Applied Ethics held in November 2010, and an earlier version of Satoru Suzuki's paper was delivered at the SOCREAL 2010: the Second International Workshop on Philosophy and Ethics of Social Reality in March 2010. Both events were organised by CAEP.

Shunzo Majima
Editor-in-Chief

Who Should Decide the Content of Professional Ethics?

Søren Holm

Manchester University, UK & University of Oslo, Norway

Abstract

This paper provides an analysis of the question: Who should decide the content of professional ethics? The main focus of the paper is on the health care professions in general and the medical profession in particular. The first section provides a brief outline of 'the canonical history' of medical professional ethics from Hippocrates to the present day and it is argued that the version of this history often assumed by the medical profession is largely false. The second section of the paper then analyses the relative legitimacy and weight of the claims to influence of different stakeholder groups in modern health care systems and argues that the normatively most important claim is that held by patients. Based on this it analyses what implications this has for the formulation and revision of professional ethics. The third and fourth section then analyses two possible set of counterclaims, one set from the professions based on a claim of epistemic superiority, 'only the wearer knows where the shoe pinches' and on a claim of the nature of a true profession; and one set from academic bioethics. It is argued that neither set of counterclaim is convincing. The final section then outlines the conclusions and briefly considers whether they are valid for professions outside of health care.

Keywords: epistemic priority, power, professional ethics, stakeholder

Introduction

Most professions have a set of ethical guidelines or rules that are binding on the members of the profession and that aim at regulating the conduct of members of the profession in relation to their clients, society at large and/or other members of the profession. These guidelines or rules are often referred to as the ethics of the profession, or professional ethics.

The purpose of the present paper is to analyse how the content of such a set of ethical guidelines or rules should be decided. Who should be involved in the process and who should have the final say? In order to set this question within a specific context the first part of the paper will give a brief overview of the development of professional ethics within the medical profession from the Hippocratic Oath to the present day.

The second part will then analyse whose interest professional ethics ought to promote and will argue for the interim conclusion that professional ethics ought to promote the interests of the clients of the profession and the interests of society. It will then explore the implications of this interim conclusion for how the process of devising professional ethics ought to proceed drawing upon stakeholder theory. One immediate objection is that everyone thinks that, at least in the health care context the clients / patients are the most important stakeholders when rules of professional ethics are being developed. And if this is true the whole paper is misguided because it argues against a non-existing straw man. But it is unfortunately not true that patients are universally recognised as the most important stakeholders or even important stakeholders. Even a brief trawl of the ethics committees of medical associations or chambers of physicians around the world (especially outside the richer countries) shows that professional ethics is still being developed in splendid isolation by physicians, often with some expert input from lawyers and philosophers. If they are lucky patients are allowed to comment on late drafts during public consultation (see, among many other sources http://www.portalmedico.org.br/novocodigo/comofoielaborado_3.asp for a history of the development of the latest Brazilian code of medical ethics and http://www.mdcnigeria.org/ for the composition of the body that is responsible for the Code of Medical Ethics in Nigeria). It is not without reason that the World Medical Association in 2010 felt it necessary to state that:

> "8.An effective and responsible system of professionally-led regulation by the medical profession in each country must not be self serving or internally protective of the profession, and the process must be fair, reasonable and sufficiently transparent to ensure this. *National Medical Associations should assist their members in understanding that self-regulation cannot only be perceived as being protective of physicians, but must maintain the safety, support and confidence of the general public as well as the honour of the profession itself.*" (World Medical Association, 2009a, my emphasis)

In the analysis the focus will be on that sub-set of professional ethics that involves the relationship between the profession, its clients and society. This means that the internal ethics of professions, i.e. how members should treat each other is outside of the scope of this paper.

The third and fourth part will then look at two sets of possible counterarguments, one from the professions and one from academic bioethics.

In the final part of the paper the conclusions will be outlined and it will be discussed whether they can be extended to other professions than the medical profession.

Setting ethical rules is not sufficient in itself. Unless professional ethics is to be only hortatory and aspirational it needs to be enforced either by the profession itself or by some external body. But questions related to enforcement are outside the scope of this paper.

The Development of Professional Ethics in the Medical Profession

The medical profession is the oldest of the health care professions. Although curing and caring as human social practices must have developed at about the same time, curing became professionalised long before caring. And from its earliest stage of professionalization the medical profession has had professional ethics in the sense outlined above.

In one version of the history of the medical profession there is an unbroken line from the Hippocratic Oath as the first expression of medical ethics to the professional ethics of the medical profession today, as it is expressed in declarations by the World Medical Association and in national codes of professional ethics. In this version of the history there has always been one profession, with one professional ethics and the interests of the patients as the focal point for that professional ethics. This 'canonical' version of the history is often accepted by medical doctors and form the basis for claims that they stand in a Hippocratic tradition and are bound by Hippocratic principles.

This version of the history is undoubtedly false. The Hippocratic school was a minority school in ancient Greek medicine and there is no unbroken line from the Hippocratic Oath to modern professional ethics (Edelstein 1943). At the most trivial level this follows from the fact that the Hippocratic Oath is sworn to Apollo, Asclepius, Hygeia and Panacea, all Greek gods or demi-Gods. No Christian, Jewish or Muslim physician can therefore swear the Oath without engaging in idolatry and in the middle ages we thus see a proliferation of versions of the Oath, 'as it may be sworn by a Christian' etc. And some even argue that there is a fundamental difference between Hippocratic and Judeo-Christian medicine (Veatch & Mason 1987). Furthermore, the medical profession as we see it today represents a confluence of several different professions (e.g. in the UK university educated physicians and apprenticed barber surgeons), and outside of Europe the medical profession has also assimilated strong indigenous healing traditions with their own sets of ethical precepts. There were for instance strong, non-Hippocratic medical professions in Japan, India and China before the advent of modern western medicine. And, finally many other values than 'the interests of the patients' have played, and still play a role in the development of professional ethics. Most important among these other values have always been the interests of the profession itself and, especially historically various religious values. Never the less the canonical version plays an important role as a foundation myth for the profession. Just as Florence Nightingale at Scutari does for nursing.

That the interests of the profession has historically played a significant role in the formulation of professional ethics can, for instance be seen in the American Medical Association's 1847 "Code of Medical Ethics". This Code was formulated at a point in time when orthodox medicine was still in public competition with many other healing practices and contains rules like the following:

> "Chapter 2, Article 1, § 3. It is derogatory to the dignity of the profession, to resort to public

> advertisements or private cards or handbills, inviting the attention of individuals affected with particular diseases –publicly offering advice and medicine to the poor gratis, or promising radical cures or to publish cases and operations in the daily prints or suffer such publications to be made ; -to invite, laymen to be present at operations,-to boast of cures and remedies,-to adduce certificates of skill and success, or to perform any other similar acts. These are the ordinary practices of empirics, and are highly reprehensible in a regular physician."
> "Chapter 1, Article 2, § 5. A patient should never weary his physician with a tedious detail of events or matters not appertaining to his disease. Even as relates to his actual symptoms, he will convey much more real information by giving clear answers to interrogatories, than by the most minute account of his own framing. Neither should he obtrude the details of his business nor the history of his family concerns."
> (American Medical Association, 1847)

In Chapter 2, Article 1, § 3 of the Code we find a number of practices condemned because they are "derogatory to the dignity of the profession" and "ordinary practices of empirics". That is, the reason given for these practices being "highly reprehensible" is not that they are ethically problematic but that they will detract from the social standing of the profession and are the kinds of practices that the competition are engaged in. And in article Chapter 1, Article 2, § 5 we find a view of the duty of patients that is clearly seen exclusively from the perspective of the profession.

And if these examples are not sufficiently convincing that the interests of the profession has played a role in the development of this code, perhaps Chapter 2, Article 4, §10 concerning how a physician should deal with questionable practice by other physicians will convince:

> "A physician who is called upon to consult, should observe the most honorable and scrupulous regard for the character and standing of the practitioner in attendance: the practice of the latter, if necessary, should be justified as far as it can be, consistently with a conscientious regard for truth, and no hint or insinuation should be thrown out, which could impair the confidence reposed in him, or affect his reputation." (American Medical Association, 1847)

Similar examples can be found in many other historical codes of professional ethics.

More modern codes of professional ethics are not as overtly giving weight to the interests of the profession (but see Kenny et al 1999 and Kipnis 2002), but these interests never the less still play a role, for instance in the way these codes conceptualise the relation between doctors and other health care professionals.

The second edition of the World Medical Association's "Medical Ethics Manual" published in 2009 has this to say about a team approach to health care:

> "Whereas relationships among physicians are governed by generally well-formulated and understood rules, relationships between physicians and other healthcare professionals are in a state of flux and there is considerable disagreement about what their respective roles should be. As noted above, many nurses, pharmacists, physiotherapists and other professionals consider themselves to be more competent in their areas of patient care than are physicians and see no reason why they should not be treated as equals to physicians. They favour a team approach to patient care in which the views of all caregivers are given equal consideration, and they consider themselves accountable to the patient, not to the physician. *Many physicians, on the other hand, feel that even if the team approach is adopted, there has to be one person in charge, and physicians are best suited for that role given their education and experience.*" (World Medical Association, 2009b, p. 90, my emphasis)

A priority for the profession is also often embedded in the process by which ethical codes and guidelines are being developed. This is almost always done internally within a professional association. There may be some input from lawyers, philosophers or theologians, but most of the input and the final say come from the profession.

Whose Interest(s) Should Professional Ethics Promote?

Who has a legitimate interest to bring to the table when the contents of professional ethics for the medical profession, or any other health care profession is to be decided? One way of answering this question is to ask who is a legitimate stakeholder in the field of activities that the rules of the professional ethics are going to govern. The concept of a stakeholder has been defined in the following way by the doyén of stakeholder theory:

> "A stakeholder in an organizations is (by definition) any group or individual who can affect or is affected by the achievement of the organization's objectives." (Freeman 1984, p. 46)

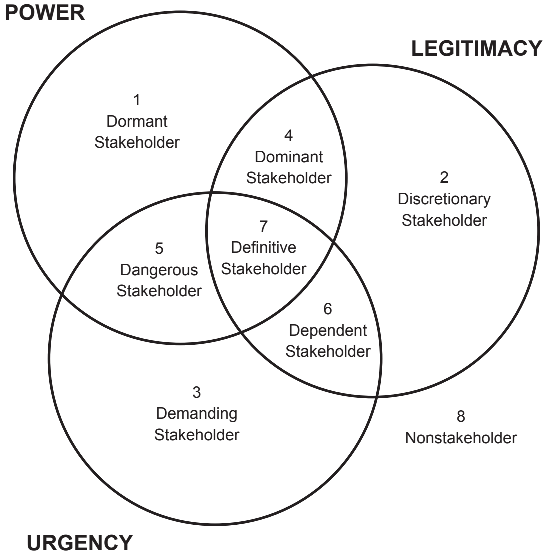
In modern health care systems this definition entails that patients, health care professionals, citizens and in certain circumstances private firms like health insurers are all legitimate stakeholders. In the commercial sector where stakeholder theory originates it is of course possible for a firm to ignore some stakeholders, even though their claims are legitimate; and our legal systems often give clear priority to one or two types of stakeholders (e.g. owners and management). But health care 'firms' are not ordinary firms because the product they deliver is not an ordinary product. The way we have chosen to organise health care clearly shows that we as citizens recognise that health care is special and socially valuable.

This means that we are working within a more normative version of stakeholder theory where we need to ask which groups of stakeholders that ought, normatively to be included in deciding on the content of professional ethics.

Mitchell et al. provides a typology of stakeholders according to three aspects of stakeholder position (Mitchell et al, 1997):

1. Power
2. Legitimacy
3. Urgency
   a. Size of stake
   b. Need to have claim dealt with urgently

This gives rise to the following graphic representation (redrawn from Mitchell et al 1997, p. 874):



Within this classification the health care professions score high on power, legitimacy and urgency; but although patients arguably score even higher on legitimacy and urgency, they often score significantly lower on power. This classifies the professions as 'definitive stakeholders' and patients as 'dependent stakeholders'.

But at a normative level patients clearly have a very strong claim to be seen as definitive stakeholders. The individual size of their stake can be a matter of life and death and the whole *raison d'etre* of the health care system is to provide services for those who have health care needs. That patients are often powerless and not included in the proces when professional ethics is formulated or revised is thus a problem. Their stake is normatively the most important and ought to be given full weight. Focusing on the patient interest does, however not entail that the patient interest must always be paramount or overriding all other interests. It only entails that it must be given its due weight. No one explicitly denies that the health care system is there for the patients, and not for its employees, the professionals. But as we have seen above this is performatively denied in the way the professions use the system, and in the present context the formulation of ethical codes to promote their own interests.

Similarly there are large differences in power between different health care professions, with the medical profession traditionally being the most powerful. In the past these differences in power might have been justifiable by the superior knowledge and training of doctors, but this is no longer the case. Other health care professions are now equally or more knowledgeable in their area of specialisation (Holm 2011).

What will giving due weight to all legitimate stakeholders and stakeholder interests mean in practice?

It has clear implications for the processes by which professional ethics is formulated and revised. We have good reason to believe that stakeholder interests are only really taken into account if those stakeholders are present in the decision making process. Relying on the professions to represent the interests of patients is, for instance not a satisfactory solution because of its inherent conflict of interest. We have to recognise that there are, potentially irresolvable conflicts between the interests of different stakeholder groups. The interests of patients do no coincide with the interests of the health care professions, and the interests of nurses not with the interest of doctors etc. And patient's cannot rely on the state to champion their interests either. As a major funder of health care, either directly or indirectly and in some systems as a major provider of health care the state must represent both the interests of (all) citizens and the interests of patients, but those interests are again in potential conflict.

Letting patients represent patient interests is, however not straightforward. One reason for the relative powerlessness of patients is that they are in an important sense external to the health care system. The professionals are much more embedded within the system. They work there every day and have clear formal and many informal lines of communication by which they can influence the policy decisions made in the system. But even though patients are, or at least ought to be the central focus of the system they are often only temporary visitors.

Another problem is that it is not clear how to represent patients. Patient organisations often represent only a specific group of patients and just as it is important to note that the interests of different professional groups might differ it is also important to note that the interests of different groups of patients might conflict. Proper involvement of patient stakeholders therefore necessarily entails having a plurality of patient representatives at the table.

Taking proper account of all legitimate stakeholders also entails a role for third party payers and for employers. Most health care is not paid directly by patients from their own resources and many health care professionals are now employees and not self-employed. This means that third party payers and employers have legitimate interests in ensuring that the rules of professional ethics, for instance promote fair allocation of resources and the accountability of employees.

Counterarguments from the Professions

A frequent general argument for professional self regulation is that only members of a profession know enough about the activities of the profession to decide on the rules that should govern the profession. The epistemic sentiment expressed in the English proverb ‘Only the wearer knows where the shoe pinches’. More formally the argument might be set out as:

> P1. Regulating a profession requires full knowledge concerning the activities of the profession
> P2. Only active members of the profession can possess such full knowledge
>
> Therefore: Only active members of the profession can regulate the profession

If sound this epistemic priority argument would entail that the analysis above is flawed. Even if there are other stakeholders, the profession’s epistemic priority entails that the views of other stakeholders should be given less or perhaps no weight at all in deciding about professional ethics. But is the epistemic priority argument sound?

Let us first note in relation to P2 that strong standpoint epistemologies claiming that only one class of epistemic agents can fully understand a specific context or activity are generally suspect. The strong standpoint claim is not equivalent to the claim that some class of epistemic agents can more easily understand a specific context or activity. It is much stronger, because it denies that a suitably motivated and cognitively capable relevant agent who is not a member of the class in question can ever reach full understanding.

If it is the case that the activities of a profession can only be fully understood by those who are active members of the profession it would have wide ranging consequences, far beyond the area of professional ethics. Let us briefly look at two of these consequences. First, the whole area of the social sciences studying the professions will be fundamentally misguided and undermined, because a social scientist who is not at the same time an active professional in the very profession he is studying will not be able to provide valid knowledge about any activity that is internal to the profession. The knowledge generated by outsiders would always be only at best partial and at worst wrong and therefore not knowledge at all. The sociology of professions could therefore, if the epistemic priority argument is sound only validly study external aspects of the profession, e.g. their societal prestige or the social background and average wages of professionals. But the sociology of the professions has given us so much more over time, including crucial insights into the internal functionings of the professions (see for instance Freidson 1970, Atkinson 1995, Helman 2001). Some of these insights have even come as a surprise for members of the professions in question (e.g. the classical studies of medical socialisation and training Becker et al 1961, Bosk 1979). This seems to imply that the epistemic priority argument is unsound. Second, the epistemic priority arguments also seems to entail that societal regulation of the activities of a profession cannot be performed by anyone who is not a member of the profession, because such outsiders could never understand the profession fully. This will completely undermine the role of regulation in modern democracy as well as the fundamental democratic ideal that in a democracy each citizen counts for one and no one counts for more than one. It will also have profound practical consequences. If banking is a profession then we will have to conclude that despite the recent fundamental failings and near collapse of our banking system, we must leave all future regulation of banking to the bankers, because they are the only ones that fully understand banking! In the banking context this is clearly a counsel of despair, since we have conclusive evidence that bankers are completely incapable of regulating themselves, partly because *they* do not fully understand banking. This again seems to point to the epistemic priority argument being unsound. A second line of criticism of the epistemic priority argument in relation to the health care professions considers the possible scope of the claim. If we accept some kind of epistemic priority, what areas of the activities of the profession plausibly fall within the scope of the claim? The most plausible part of the claim covers those areas of activity where only the profession is engaged, e.g. the internal relationships between members of the profession. What kind of deference, if any should junior members of the profession for instance pay to senior members? But the claim to epistemic priority is logically undercut as soon as a member of another profession or a non-professional is involved in the interaction. Neither doctors nor nurses can plausibly claim epistemic priority concerning their inter-professional interactions. Both are in the situation and must be epistemically equivalent. More important for the present discussion is that some of the ethically most important interactions of the health care professions are with patients, and that it seems very strange for the professions to claim epistemic priority concerning these interactions. Why should we believe that doctors appreciate the full ethical import of these interactions better than patients? Let us, for instance take the example of abortion. We may have some reason to accept the claim that obstetricians are in a better position to understand how it is to perform an abortion than many other people, but we can at the same time deny the claim that obstetricians are in a better position to understand how it is to need an abortion or to have one performed! Obstetricians thus have important knowledge that should be part of discussions about how we should regulate abortion, but so do pregnant and non-pregnant women and many other groups. Obstetricians cannot, based on epistemic priority claim that abortion should be regulated by their professional rules into which only they have any input.

Finally it is worth noting that if the epistemic priority claim is sound, then it completely undercuts any claim of the health care professions to speak on behalf of, or represent their patients. If only patients can fully understand how it is to be a patient, then even the most well meaning doctor is misguided if the thinks he can represent them and their experiences. The doctor may of course know more about the patient’s disease than the patient does in the abstract, but unless the doctor has also himself had that disease the epistemic priority claim entails that he cannot fully understand how it is to be a patient (see also Holm 2005).

Based on all of the considerations taken together the epistemic priority claim for the professions either fails, or becomes completely implausible.

A second possible claim is that a profession is only a true profession if it is fully self-governing, i.e. that it itself both sets and enforces the rules governing the practice of its members. If others are involved in setting these rules the profession is no longer a true profession. What is the status of this claim?

Let us first note that it has the implication that the medical profession is not a true profession in many countries in the world and that this situation has obtained for a very long time. There are many countries where the medical profession is not self-governing in the sense referred to above.

But perhaps more fundamentally we can ask whether we should or can allow true professions in modern society? Are true professions compatible with democracy? Let us imagine a situation where a country decides that a certain health care service A should be available in the health care system, but where the medical profession refuses to provide this service and decide that no doctor should perform A. There are clearly circumstances where the profession will be justified in its stance, for instance if A is grossly immoral, but in that case the justification does not rely on the profession being a true profession and self-governing. If A is grossly immoral no one should provide it, whether or not they are professionals. But if the justification provided by the profession for not providing the service is not of this general nature it is unclear why society should allow the profession to override democratic decision making.

The Counterargument from Academic Bioethics

The view that all stakeholders must be involved in formulation and revision of professional ethics for the health care professions can and has also been criticised from within academic bioethics.

This criticism asks why we need stakeholder involvement or the involvement of medical sociology if we have a satisfactory general ethical framework? (Herrera 2008) Based on a satisfactory general ethical framework we ought to be able to develop a satisfactory professional ethics without any specific stakeholder involvement. From a consequentialist point of view we could, for instance conceptualise professional ethics as a particular kind of rules aimed at maximising good consequences in a specific sphere of practice. The task would then be to develop the optimal set of rules from this perspective. This requires knowledge about the sphere of practice, but not necessarily any involvement of stakeholders as stakeholders.

There are at least two possible answers to this critique. The first answer accepts the critique in principle, but points to the fact that the conclusion has no practical relevance. Even if we accept that we could develop a satisfactory professional ethics from a satisfactory general ethical framework, we have the problem that we have not, so far been able to agree on a satisfactory ethical framework. I may believe that I have one, and you may believe that you have one, but if I am a libertarian and you are a consequentialist we are still not in a position to derive professional ethics from a general framework.

The second answer to the critique also accepts the critique in principle, but points out that applying a general ethical framework to a specific context is not a simple matter. It is now more than 20 years ago that Caplan provided his devastating critique of the 'engineering model' of applied ethics (Caplan 1987), i.e. the view that ethical principles can be applied in a way similar to engineering principles and formula. And Caplan's arguments apply equally well to the formulation of rules of professional ethics. It does not follow directly from this that we need to involve stakeholders, but on the plausible assumption that we need to involve someone who are knowledgeable about the specifics of the context and the interests of stakeholders it does follow that we need to involve more than moral philosophers. Who should we then involve? We could possibly get this detailed knowledge from medical sociologists and other social and political scientists. But, given that stakeholders have legitimate claims and interests and given that they may understandably see a process where they are not involved as less legitimate than one where they are, it seems at least prudent (and probably ethically required) to involve stakeholders.

Conclusion

This paper has argued for the view that the formulation of professional codes of ethics for the health care professions cannot be left to the professions themselves. The professions have multiple, clear potential conflicts of interest and they cannot plausibly claim to represent the main stakeholder in health care, i.e. the patients. All legitimate stakeholders must be involved in the processes of formulating and revising professional ethics and the degree of involvement and the weight in decision making should reflect the normative importance of their stake.

Are these conclusions valid for other professions as well? This depends on a number of factors, primarily on what the relation is between the profession and those people it serves and acts upon. For professions that are like medicine in the sense that the members of the profession acts for or on behalf of their clients very similar conclusions follow. But there are professions where this relationship is more complicated. A brief consideration of the profession of police officer makes this point stand out clearly. The police service acts of behalf of all law abiding citizens, but part of its clients are clients exactly because they are not law abiding and because it is the duty of the police service to act against their interests. At least if we can assume that it is not in the interest of the criminal to be apprehended, arrested, sentenced and punished. This means that although criminals are what we could call nominal stakeholders in relation to the ethics of policing according to Freeman's definition of a stakeholder, they are not normative stakeholders. They do not have a strong, legitimate claim to have their criminal stake protected or to be part of the process of formulating police ethics.

References

American Medical Association (1847). *Code of Medical Ethics*. Philadelphia: American Medical Association. Available in facsimile at http://www.ama-assn.org/resources/doc/ethics/1847code.pdf

Atkinson, P.A. (1995). *Medical Talk and Medical Work*. San Fransisco, CA: Sage Publication.

Becker, H.S., Geer, B., Hughes, E.C. & Strauss, A. (1961) *Boys in White: Student Culture in Medical School*. Chicago: University of Chicago Press.

Bosk, C.L. (1979). *Forgive and Remember: Managing Medical Failure*. Chicago: University of Chicago Press.

Caplan, A.L. (1987). Can applied ethics be effective in health care and should it strive to be?. In: Ackerman TF, Graber GC, Reynolds CH, Thomasma DC (eds.) *Clinical Medical Ethics – Exploration and Assessment*. Lanham, MD: University Press of America. pp. 131-143.

Edelstein, L. (1943) *The Hippocratic Oath: Text, Translation, and Interpretation*. Baltimore: Johns Hopkins University Press.

Freeman, R.E. (1984). *Strategic Management – A Stakeholder Approach*. London: Pitman.

Freidson, E. (1970). *Profession of Medicine*. Chicago: Chicago University Press.

Helman, C.G. (2001). *Culture, Health and Illness*. London: Arnold.

Herrera, C (2008) Is it time for bioethics to go empirical? *Bioethics* 22(3): 137- 146.

Holm, S. (2005).Justifying patient self-management – evidence based medicine or the primacy of the first person perspective. *Medicine, Health Care and Philosophy* 8(2): 159-164.

Holm, S. (2011). Final responsibility for treatment choice – the proper role of medical doctors?. *Health Expectations* 14(2): 201-209.

Kenny, N., Weijer, C. & Baylis F. (1999) Voting ourselves rights: a critique of the Canadian Medical Association Charter for Physicians. *CMAJ* 161(4):399-400.

Kipnis, K. (2002). Ethical Competency and the Profession of Medicine. *Virtual Mentor,* 4(8) http://virtualmentor.ama-assn.org/2002/08/code1-0208.html

Mitchell, R.K., Agle, B.R. & Wood, D.J. (1997)Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of Who and What Really Counts. *The Academy of Management Review*, 22(4): 853-886.

Veatch, R.M. & Mason, C.D. (1987) Hippocratic vs. Judeo-

Christian Medical Ethics: Principles in Conflict. *The Journal of Religious Ethics,*15(1): 86-105.

World Medical Association. (2009a) *WMA Declaration of Madrid on Professionally-led Regulation*. Ferney-Voltaire: WMA.

World Medical Association. (2009b) *Medical Ethics Manual* (2. Ed.). Ferney-Voltaire: WMA.

Neuroimaging in the Courts of Law

Bernard Baertschi

University of Geneva, Switzerland

Abstract

Lie detection has recently become a topic of discussion once more. Courts of law were interested in it for a long time, but the unreliability of the polygraph prevented any serious use of it. Now a new technology of mind-reading has been developed, using different devices that are deemed to be able to detect deception, in particular Functional Magnetic Resonance Imaging (fMRI). Is fMRI more reliable than the polygraph? It meets at least with various kinds of obstacles: technical, methodological, conceptual and legal. Technical obstacles are linked with the state of the technique, methodological ones with epistemological difficulties, conceptual ones with problems tied to what lying consists of, and legal ones with the effects of brain imaging on lawsuits. I examine several of these and conclude that at present mind-reading using fMRI is not ready for use in the courts. The obstacles examined may not be insuperable, but a lot more research is needed.

Keywords: brain imaging, mind-reading, lie detection, courts

1. Introduction

For a long time, human beings have tried to decipher the mental states of their fellows without relying on what they say or might say about themselves. There is very good reason for this: in particular, if we could read the minds of other people, some drawbacks in our social and moral life could be avoided. Importantly, we could detect liars and cheats, a very crucial matter to assess others and to improve human cooperation (Cosmides and Tooby, 2008). In this context, it is not surprising that judges are interested in mind reading. In the United States, the polygraph has been used (and is still used) in particular States, but it has never been acknowledged as a reliable tool. However, for some time now a new tool has been available – neuroimaging. Is it more reliable? This is the question that I will address in my paper.

Several devices have been developed to look non-invasively inside the human body beyond traditional X-rays. They are often referred to by the generic term 'scanners'. The more important are Electroencephalography (EEG), Positron emission tomography (PET) and Magnetic resonance imaging (MRI). Physicians are using them for a lot of medical purposes, but as scanners can also 'look' under the skull they can be used to look into the brain and, through it, into the mind. Functional magnetic resonance imaging (fMRI) is now widely used by neuroscientists and psychologists to study the human mind.

In the courts, two subjects are linked to brain imaging; the detection of liars and the insanity defence. I give 'insanity defence' a wide meaning: it includes every claim to mitigate responsibility on the basis of some structural or functional brain abnormality. It is a wider subject than the first, but, in this paper, I will speak exclusively of the detection of liars, for two reasons. First, it has been hotly debated recently in the United States, and on many grounds we can expect that forensic use of scanners will spread in other countries. Second, the questions raised by the use of neuroimaging for detecting liars are relevant for the general reliability of brain imaging in neuroscience.

In the United States, some companies already offer lie detection through EEG (Brain Fingerprinting) or fMRI (Cephos, No Lie MRI). Walter Sinnott-Armstrong and colleagues claim that 'EEG data were admitted as evidence against lying in 2001 by Iowa District Court Judge Tim O'Grady in the case of Terry Harrington' (2008, 360). Terry Harrington was convicted of murder in 1978, and freed in 2001. But if the EEG technique was accepted by the court, contrary to what is very often said, it was not tested on this occasion, because while the district court has been ordered to make a new trial for reasons unrelated to lie detection, it gave up: 'The local prosecutors declined to pursue the case and Harrington was freed' (Wolpe et al., 2005, 43). In the last few years, it has been successfully introduced several times in suits, but more recent attempts, in 2010, have failed (Saenz, 2010). In India, the technology is now being considered for use (Racine, 2010, 3).

How is lying detected by these devices? Two techniques are used: the Control question test (CQT) and the Guilty knowledge test (GKT). As Paul Wolpe and colleagues explain (2005, 40), in CQT the subject is asked to answer three kinds of yes-no questions: relevant questions (e.g. 'Did you kill your wife?'), control questions (e.g. 'Did you ever steal something?') and irrelevant questions (e.g. 'Are you sitting in a chair?'). Everybody is expected to react more strongly to control than to relevant questions, except the culprit. Therefore, if the defendant denies having killed his wife but reacts more strongly to relevant questions than to control ones, then it is a clue that he is lying.

In GKT too, a series of questions are asked, some relevant, some irrelevant or neutral relative to knowledge that only the guilty person could possess concerning the place, time and details of the crime. 'For example, in a crime investigation involving a stolen red car, a sequence of questions could be: "Was the car yellow? Was the car red? Was the car green?" The questions are chosen so that subjects with knowledge of the crime (but not other individuals) would have an amplified physiological response to the relevant question – that the car was red – which is dubbed "guilty knowledge"' (Wolpe et al., 2005, 40). GKT was the technique used in Terry Harrington's case.

As we can see, lie detection is more direct with CQT, but both techniques can be used to test the truth. The underlying hypothesis is that brain activity changes in a controlled manner when somebody tells a lie. Is this true? To answer this question, it is necessary to review thoroughly the tools and techniques used.

2. Some Technical Obstacles

I will limit my investigation to fMRI, because it is the device that has been mostly scrutinised in peer-review papers. In order to use fMRI, be it for experimental or for forensic purposes, the task or procedure must first be calibrated to be adapted to the subject or the defendant. The subject can then engage in the mental task under investigation (e.g. lying). This still cannot be recorded directly: as we have seen, such a measurement always rests on a comparison of two tasks (or more), as Eric Racine and colleagues explain:

> An assessment of brain activity during the mental process of interest then resides in statistical comparisons across the entire brain (separated into cubic elements or voxels), of the signal level between the control task and the experimental task. Those brain regions, or voxels, which meet statistical significance at a set threshold level, differentiate areas where the signal is statistically greater in the active versus control task. These regions are labeled in a colour coded map to represent active brain areas, and are generally overlaid onto a structural image. (2010, 246)

Briefly said, coloured spots in brain pictures represent variations in blood oxygen, variations that are recorded only if they are above a threshold and characteristic for the experimental task investigated.

As we can see, using fMRI is not an easy matter: brain imaging has nothing to do with photographs. Therefore, it is not surprising that the technique encounters some specific obstacles. I will examine three of them: (i) replicability of results, (ii) BOLD factors and cognition, and (iii) method of subtraction. There exist other technical obstacles, i.e., obstacles arising from the limitations of the technique used, the most well-known being that brain activity is much faster that fMRI measurements, but I will not dwell on them.

By *replicability of results*, I do not refer to the applicability of laboratory experiments to real-world scenarios. This applicability is a real problem, but I will ponder on it when I will attend to legal obstacles. What I mean by replicability of results is the problem created by what has been called 'uniqueness of neural signature' (VanMeter, 2010, 237). Every brain has its singularity and is different from other brains; consequently, there exists an individual variability that could prevent the acquisition of universal inductive knowledge on lies applicable to all individuals. As we can see, this obstacle is technical *and* methodological or epistemological; therefore it is not certain that technical progress will be able to overcome it.

This variability is individual *and* general: we have observed some structural differences between the brain of psychopaths and of normal persons; the effect of these differences could be that the neural signature of lies is not the same in a psychopath and in a normal person.

The problem of replicability is particularly important for brain imaging, because fMRI pictures representative of a certain mental task and used as standards report not individual brains, but group-average brains. Sinnott-Armstrong and colleagues notice: 'Individual functional profiles can vary so much that it is not unusual for most individuals to differ from the group average' (2008, 362). It means that it will sometimes happen that the brain image of a particular liar will be sufficiently different from the 'typical' brain image of a liar, that he will be considered as a truth teller. The reverse is possible, too.

MRI measures *BOLD factors* – i.e., blood oxygenation level dependent factors – in short, the amount of oxygen present in blood. But, as Valerie Hardcastle and Matthew Stewart say: BOLD factors are not the same thing as cognition (2009, 187); it is not even identical with brain activity. Between BOLD factors and cognition, there are two intermediate levels; brain activity and coloured voxels (the 3D pixels visible on brain pictures). To pass from one level to another requires interpretation – some authors even speak of 'manipulation' –, and interpretations can lead us astray.

Interpretation of data is difficult, but obtaining relevant data is not easier. To gather the data, we must first set the threshold of the MRI device in order to avoid false positives *and* obtain useful results. If the threshold is too low, we are overwhelmed by a wealth of uninterpretable voxels or false positives, like the dead salmon perceiving human feelings! Bennett put a dead salmon in an MRI machine and observed some activity characteristic for the perception of emotions in third parties in the fish's brain, although it is absurd to claim that a dead salmon can perceive anything. However the problem cannot be solved simply in setting the threshold much higher, because, as Bennett points out: 'We could set our threshold so high that we have no false positives, but we have no legitimate results' (Madrigal, 2009.2, 2). A threshold must be determined, but there is no easy way to do it.

When the threshold is set, the usual way to obtain relevant data is the *method of subtraction*. As we already know, the subject performs two tasks, the data obtained are subtracted and the remaining data is specific for the task the researcher wants to investigate. Marcus Raichle illustrates this method with an example: 'For example, to "isolate" areas of the brain concerned with reading words aloud, one might select as the control task passively viewing words. Having eliminated areas of the brain concerned with visual word perception, the resulting 'difference image' would contain only those areas concerned with reading aloud' (2009, 5).

Comparing results is ubiquitous in brain imaging. We compare because we are unable to observe directly. But it is not without dangers, as Adina Roskies states: 'The very same raw data from the main task can give rise to a very different image depending on which comparison task is employed in the analysis, and very different tasks can give rise to very similar images if appropriate comparison tasks are chosen. Neuroimaging, unlike photography, is essentially contrastive' (Roskies, 2007, 870). The result of a subtraction depends on the numbers used in the calculation; if one number changes, the result will of course not be the same.

3. Some Methodological Obstacles

Let us imagine that the technical obstacles are overridden, thanks to technological progress (a super-fMRI is available), and suppose that we observe constant correlations between brain patterns and mental tasks, delivering replicable results. Would all the obstacles be suppressed and could the judges confidently rely on brain imaging? Unfortunately, it will not be the case, because there will still be obstacles of a more pervasive nature. These obstacles are methodological or epistemological and three of them are prominent in my mind: (i) correlation and explanation, (ii) correlation and causation, (iii) reverse inferences.

Correlation is not explanation. From the fact that *A* correlates with *B*, it does not follow that *A* explains *B*. We observe correlations between brain events (BOLD factors) and mental events, and even constant correlations. Therefore we are tempted to conclude that brain events explain mental events. For instance, from the fact that amygdala activation is strongly correlated with anger and fear, we jump to the conclusion that fear and anger are explained by amygdala activity. This conclusion is too hasty for two reasons. First, it draws on a metaphysical thesis (like brain/mind identity, epiphenomenalism or supervenience of mind on brain) without arguing in its favour. Second, and more importantly – because we have in fact a lot of independent arguments in favour of some materialist metaphysical thesis (Kim, 1998) –, other neuroscientific explanations are possible, as Ellis states:

> It remained for a few maverick neuroscientists, such as Panksepp, to keep insisting that amygdala activation was not the substrate of anger, but instead might correlate with anger only because it played an important role in learning and remembering which stimuli should elicit anger. Panksepp's view is that the periaqueductal grey area deep in the subcortex is the most crucial part of a complex anger circuit involving many brain areas other than the amygdala. (2010, 69)

Explanations need more than correlations.

Correlation is not causation. We have known for long time that succession is not causation (*pace* Hume); constant succession is a kind of correlation, i.e., a relation that is not law-like, contrary to causality.

Therefore, if constant succession is not causation, it is because correlation is not causation. It is easy to show why. If *A* correlates with *B*, it does not follow that *A* causes *B*. *A* and *B* can for instance be the joint effect of *C*. The hypothesis of Panksepp could illustrate this argument, too. In his view, amygdala activation is not the cause of fear, but both are something like joint effects of events taking place in the periaqueductal grey area deep in the subcortex.

Technical obstacles did show that we do not observe a bi-univocal correspondence (a one-to-one relation) between types of brain events or brain pictures and types of mental events. The methodological obstacles examined so far add epistemological reasons to the technical ones. It does not mean that it is logically impossible to draw a bi-univocal correspondence between both types of events: when we discover laws in nature we sometimes succeed at that precisely; but for now we are unable to ascertain that it will be possible for the brain. The difficulty is made worse by the fact that mental events are private and therefore hidden; consequently, to know them, we are obliged to make *reverse inferences*. What does it mean? In establishing correlations between the mental and the cerebral, we give the subject some mental task to perform and we observe the correlated activation pattern through fMRI. When we want to know which mental task another subject is performing, we must predict it on the basis on the brain pattern observed. This prediction consists in a reverse inference (inference from brain to mind, grounded on inferences from mind to brain), but such an inference is not reliable as a bi-univocal correspondence between brain and mind has not been established. Kamila Sip and colleagues apply it to lie detection:

> When using functional magnetic resonance imaging (fMRI) (or any other physiological measurement) to detect deception, we are confronted with the problem of making reverse inferences about cognitive processes from patterns of brain activity. Even if deception reliably activates a certain brain region, we cannot logically conclude that, if that brain region is activated, deception is taking place. (2007, 50)

4. Some Conceptual Obstacles

The technical and methodological obstacles I have mentioned and examined concern brain imaging using fMRI in general, not only its forensic use. If neuroscientists and neuropsychologists are confronted with these difficulties, judges and lawyers will be, too. In a sense, the problem is greater in the courts, because lawyers are not so well aware of the scope and limits of the technology. In another sense, however, the problem is perhaps more manageable: in the courts, we are interested in lies and in nothing else, a very narrow topic. Consequently it would not be necessary to identify the correct brain explanation of lying or its real cerebral cause to detect lies: correlations like those observed in CQT and GKT tests could suffice. Would it not be possible then to reach an agreement concerning the neural signature of lies, even if in certain respects this signature is manifold?

To reach an affirmative answer to this question, two conceptual difficulties must still be solved concerning: (i) the nature of lying, and (ii) the importance of intentional context.

We wish to detect liars, but what does it mean, to lie? What is *the nature of lying*? At first sight, the answer seems unproblematic: to lie is to conceal the truth. But there are many ways to conceal the truth. What is the difference between a spontaneous and a prepared lie? A temptation to lie and an effective lie? An exaggeration and an omission? Another precision is in order: to detect someone as a liar presupposes that the liar knows that he is lying; if he does not know, his brain will not show the right activation pattern. What to do then with self-denial and self-deception? Often, an ingrained liar is persuaded that he is sincere, and if he thinks he is sincere, is he lying? As for everyday life, for the law too there exist several kinds of lies. Besides, some lies are innocent lies and an important ingredient in our life. Marcel Proust said that without lies, life would be impossible to live (Proust, 1999, 2063); for instance, we frequently lie in order to protect the peace of mind of ourselves and of our family, and to preserve good relationships with our fellows. What does neuroimaging have to say concerning this diversity of lies? Nothing at all, says Justice Jed Rakoff:

> A little white lie is altogether different, in the eyes of the law and of common sense, from an intentional scheme to defraud. Nothing in the brain-scan approach to lie detection even attempts to make such distinctions. And what might a brain scan be predicted to show in the case of a lie by omission […]? In my experience, these are the most common kinds of lies in court. (2009, 44-45)

This presents a very bad situation: how can we detect liars if we do not know what we are looking for? Such a problem is not linked with the technique used or with the methodology, it is conceptual or semantic: we must know what it means to lie if we want to test lying with fMRI or any other tool. Nevertheless, neuroscientists and neuroethicists are aware of the problem, and they have proposed a precise concept of lie to be tested. Kamila Sip and colleagues summarise it in the following way: ‘A useful characterisation is provided by Vrij, who defined deception as follows: "A deliberate attempt, without forewarning, to create in another a belief which the communicator considers to be untrue"' (2007, 48). In their comment, the authors underline that two things are important in this definition. First, the crucial point is not the truth value of what is said, but the intentional and deliberate attempt to deceive (if it not the case that *X* and I believe that *X*, then I do not lie when I say that *X*; I just make a mistake – if my interlocutor listens badly to what I say and understands not-*X*, I do not lie either). This point was established long ago by Augustine in his seminal treatise on lying (*De Mendacio*). Second, the liar is not instructed to lie (the lie occurs 'without forewarning'), but he decides freely and voluntary to deceive.

This last point prompts a new methodological problem because, in the vast majority of studies, subjects are instructed to lie, therefore they know that they ought to lie. Consequently, what fMRI detects has little to do with authentic lies, as Nancy Kanwisher states: 'Making a false response when you are instructed to do so isn't a lie, and it's not deception. It's simply doing what you are told' (2009, 12). The methodological problem is the following: to fulfil a task, the subject must know what this task consists in, therefore he must be instructed to lie if lying should be measured. But to be instructed to lie is not to lie. Consequently it is impossible to test authentic lies with fMRI. Nevertheless, the situation is different in the courts: defendants are not instructed to lie, therefore it seems that this methodological problem could be bypassed. Unfortunately, this is not the case, because lie detection standards must be established before this technique could be employed in the courts, and these standards must be set in experiments, i.e., in situations where authentic lies cannot be tested.

Perhaps it would be possible to take an indirect path. When someone is sincere or lies intentionally, it necessarily has an effect on his cognitive and emotional states. In particular, it is more difficult to lie than to tell the truth, because the liar must make an effort to invent some falsehood, and even some plausible falsehood. In consequence he will be more stressed, and sometimes embarrassed or anxious. The polygraph was invented to measure the physiological correlates of those states; in a parallel manner, fMRI could be used to measure their brain correlates. As Victoria Holderied-Milis says: 'Even when liars manage to stay undetected, their integrity is damaged nevertheless. Because of the difference between what they hold to be true and what they articulate, they endure an internal tension, which also has to be hidden from their fellows' (2010, 110-1). Many psychologists agree and think that the tension evinced in lying denotes an inhibition on telling the truth. Now, it is possible to test inhibitions with fMRI, in particular with the help of the Sternberg Proactive Interference Paradigm. Elizabeth Phelps explains it in the following way:

> In a typical version of this paradigm, a participant is shown a set of stimuli and told to remember it. For example, the set might include three letters, such as B, D, F. After a short delay the participant is presented a letter and asked, 'Was this letter in the target set?' If the letter is D, the participant should answer 'yes.' In the next trial the participant is given another target set, such as K, E, H. At this point, if the participant is shown the letter P, she or he should say 'no.' If the participant is shown the letter B, the correct answer is also 'no.' However, for most participants it will take longer to correctly respond 'no' to B than P. This is because B was a member of the immediately preceding target set (B, D, F), but it is not a member of the current target set (K, E, H). […] To correctly respond 'no' to B on the current trial requires the participant to inhibit this potential 'yes' response and focus only on the current target set. (2009, 19)

Brain imaging has shown that the inferior frontal gyrus plays an important role in this type of inhibition, and even if the above experiment has nothing to do with lying, but with our access to truth, it is thought that this region could be involved in lying.

Consequently, there seems to be a possible way out of this new methodological problem, if we adopt the following psychological thesis: lying requires inhibition of telling the truth and this inhibition can be correctly detected. The inhibition hypothesis looks plausible, but is it true? In other words, can we confidently say that all or most intentional lies are linked with such an inhibition? If that is the case, it would mean that the liar would want spontaneously to tell the truth, but he blocks this reaction, which requires an effort. Unfortunately, a recent study casts some doubts on this hypothesis.

Joshua Greene and Joseph Paxton have conducted a study on moral decision. Subjects were instructed to predict the outcomes of coin-flips. When their predictions were correct, they gained some money, but when they were not correct they were financially punished. Subjects made their prediction privately and checked themselves the result, without any supervision. Therefore it was possible for them to report falsely, i.e., to cheat. As the coin-flips were randomly made by a computer, the probability of an accurate prediction was 0.5. Therefore all subjects who claimed that they had been able to predict correctly the outcomes of coin-flips with a probability higher than 0.7 made a false report (i.e., they lied) on several occasions.

Did they have to refrain from telling the truth, i.e., from reporting the correct result each time they did not? No. The additional brain activity, deemed relevant for deception, was actually observed in fMRI, but especially in dishonest subjects (i.e. subjects who claimed that they had been able to predict with a probability higher than 0.7) who *renounced* a dishonest gain, that is, those who made an honest report of their failure to predict accurately the outcome of the flip.

Therefore, even if lying implies an effort not to tell the truth for occasional liars, telling the truth, that is not to lie, implies a more important effort for usual liars. The authors conclude:

> We find that control network activity is most robustly associated, not with lying per se, but with the limited honesty of individuals who are willing to lie in the present context. It is unlikely that control network activity associated with limited honesty is related to overcoming a default honesty response because such responses are themselves honest. (2009, 12509)

This study forces us to conclude that lying is not regularly or *per se* correlated with an inhibition on telling the truth.

This study is particularly important for the reason that the courts are frequently confronted with dishonest individuals. This drives us to *the importance of the intentional context*, the second conceptual obstacle. Discriminating between the brain reactions of honest and dishonest people already takes intentional context into account. This context is wider, however, and consists in the desires, beliefs and mental attitudes of the individuals who are lying. Kamila Sip and colleagues underline this situation with the striking example of psychopaths: 'The lack of emotional response observed in psychopaths, when they tell a lie, stems from their attitude to the victim of the deception. They have no empathy for the potential suffering that their actions might cause.' (2007, 52). This lack of emotional response is the psychological counterpart of the structural brain differences I have already alluded to, so that the neural signature of lies should be different in a psychopath and in a normal person. This is an extreme example, but even in the life of normal individuals, the intentional context has obviously an impact on the way we behave, verbally and non-verbally. A law court is a peculiar place, generating its own intentional context (think of a defendant who tries to exonerate oneself). With this remark, we come to the legal obstacles.

5. Some Legal Obstacles

Beyond the use of a common device, fMRI, and the several common obstacles it encounters, the domain where lie detection takes place (i.e. the court) generates its own problems: the domain of experimentation is not identical with the domain of trials. Trials obey specific rules: legal and procedural ones. These rules create some new obstacles for the detection of liars by means of technological devices like fMRI. I will examine four of them: (i) a trial is not an experiment, (ii) effects on juries, (iii) the role of juries, and (iv) the importance of behaviour.

A trial is not an experiment, therefore lying in a trial is not the same thing as lying in an experiment. More generally, it is doubtful that the results obtained in experiments are transferable to a litigious environment. There are two reasons for that. First, as we have already seen, lies in an experimental setting are either not authentic lies or no lies at all. Second, even if they were authentic intentional lies, the emotional, intentional and institutional contexts are so different that the brain images will be unreliable. The persons investigated are diverse, too; students on one side, defendants on the other.

The difference in context has another impact. Researchers claim that they have been able to identify lies with an accuracy ranging from 76 to 90% (Madrigal, 2009.1, 2). Commercial enterprises give even more favourable figures, ranging from 90 to 95% (look at their websites, the numbers are changing with time). This represents good results in the context of experiments, but not so good in a trial; 76% accuracy means 24% error, that is, almost one quarter. If 90% is better, it leaves nonetheless 10% error. Of course, not every error would result in a miscarriage of justice, because fMRI data would be only one piece of evidence among many. However, the nature of brain pictures and their psychological effects are a source of concern; this constitutes the second legal obstacle.

Brain pictures could have a devastating *effect on juries* in that they could influence their minds widely beyond the evidence the images can afford. This has been studied by Gurley and Marcus not in cases of lie detection, but of the insanity defence or NGRI (not guilty by reason of insanity), a study reported by Sinnott-Armstrong and colleagues in these terms:

> Gurley and Marcus (2008) found that the percentage of subjects who found the defendant NGRI after reading expert testimony on mental disorder (psychopathy/psychosis) was higher when accompanied by a brain image (19/37%), by testimony about traumatic brain injury (27/43%), or by both (44/50%) than when subjects received neither (11/22%). Thus, the introduction of both testimony about traumatic brain injury and images of brain damage increased the NGRI rate from 11% to 44% in the case of psychopathy. That is a big effect, so brain images and neuroscience do seem to affect legal decisions. (2008, 369-70)

However, as the authors comment, it does not prove that jurors are unduly influenced; maybe MRI gives them the information they need to decide correctly.

This optimistic stance is nevertheless not warranted, in the sense that we know that brain images have a strong influence on people (Racine, Bar-Illan and Illes, 2005). Several studies have demonstrated that when a brain picture is added to an argument or a thesis, it appears more convincing. The same is true if brain information is added. For instance, Deena Skolnick Weisberg and colleagues (2008) have shown that, for students, the addition of neuroscience information moderately increases their confidence in good explanations, but worse it blocks lucidity and transforms bad explanations into good ones in their minds. The authority of pictures and of science can be misleading for an uninformed public. As jurors belong to the public, they will be prone to be misled, as Paul Wolpe and colleagues state: 'Brain scan images might influence juries even when the images add no reliable or additional information to the case' (2005, 47).

This conclusion is corroborated by a very recent study led by David McCabe and colleagues, the first made on fMRI and lie detection: 'Results showed that presenting fMRI evidence suggesting the defendant was lying about having committed a crime was more influential than any other evidence condition' (2011, 574). This is not surprising, but a source of true concern. Nevertheless, there exists some hope: when subjects are informed of the limitations of fMRI, the confidence aroused by it decreases to the same level as that of other evidence conditions.

For some authors, brain imaging can have a second bad effect on juries. The *role of juries* is to assess what witnesses say and to weigh the arguments presented. fMRI would change that and could nullify the role of juries. With this argument, a defence attorney has successfully pleaded against the admission of brain imaging in a trial: 'Defence attorney Jessica Cortes of the firm Davis and Gilbert won her motion to exclude the evidence without getting into the science behind brain scans. Juries are supposed to decide the credibility of the witness, she argued, and fMRI lie detection, even if it could be proven completely accurate, infringes on that right' (Madrigal, 2010, 1). A success for the attorney, but grounded on a weak argument: scientific data has been admitted by courts for a long time. Fingerprints and DNA are daily invoked as evidence, and without any substantial change in the role of juries. The admission of fMRI data would probably not introduce any further change. And even if it did, would it be a great loss for justice if the role of juries was modified or even nullified? In many countries, juries have lost importance, due to the growing complexity of cases. This could well be an improvement, but that is another debate. For the time being, we have a lot of other reasons to be wary about the use of brain imaging in the courts. This will be my conclusion below. But I must still address one last legal obstacle: the *importance of behaviour*.

When we discuss technical matters, we often tend to forget what is really at stake. fMRI could be a useful tool or a useless tool, but it will remain just a *tool* for the institution of justice. The aim of this institution is to assess behaviour: courts judge behaviour, not brains. Therefore, a structural or functional brain picture becomes relevant only if a person has broken the law, or is accused of it – i.e., to have behaved unlawfully. Sinnott-Armstrong and colleagues emphasise: 'What matters to law is not brain function but behaviour, and abnormal brain function does not necessarily make abnormal behaviour likely' (2008, 364). Stephen Morse had already insisted on that point when the question of the death penalty for teenagers was debated in the United States. Commenting on *Roper v. Simmons*, he said: 'If the behavioural differences between adolescents and adults are slight, it would not matter if their brains were quite different. Similarly, if the behavioural differences are sufficient for moral and constitutional differential treatment, then it would not matter if the brains were essentially indistinguishable' (2006, 48).

The impact of this last obstacle spreads widely beyond the question of the use of brain imaging in the courts, but it is useful to recall it in order that fMRI stays in its proper place, if ever it is accepted as evidence by courts.

6. Conclusion

In this paper, I have listed and reviewed a large number of obstacles to the use of neuroimaging in the courts. Other obstacles exist, but I have discussed the ones I find philosophically the more interesting. What is the result of this examination? It appears to be widely negative: fMRI and other devices to detect liars should not be used in the courts, because the present obstacles are huge. But I have also said that fMRI should stay in its proper place, if ever it is accepted as evidence by courts. Does fMRI have a proper place? I consider that it has one, or rather that it will have one when its more important obstacles are overcome. I particularly think of technical and conceptual ones. Methodological and legal obstacles are of a more general nature: every piece of scientific evidence invoked in a lawsuit meets with them. Therefore they are reasons to be circumspect, but not to exclude scientific evidence from the courts. Technical obstacles are limitations that could be overcome in the future, at least sufficiently: if measurements become more precise in space and in time, difficulties with the interpretation of BOLD factors and with subtraction method could be lessened, and we will perhaps have a better way to manage the uniqueness of the neural signature. Conceptual obstacles will require conceptual and empirical work to be overcome. We must make efforts to develop a more precise psychological theory of lie and deception, a theory that will inspire neuroimaging experiments. This is not out of our reach at all, but it will take time.

The problem is finally not with the use of fMRI or other devices for lie detection in the courts *per se*, but with a premature adoption of an existing device. And if there is an actual risk in a premature adoption, it is because the brain imaging technology is not well understood by certain of the interested parties. I hope that this paper will throw some light on this.

Acknowledgments

I wish to thank Patrik Vuilleumier for his helpful comments. An earlier version of this paper was presented at the 5th International Conference on Applied Ethics, at Hokkaido Universty, on 6 November 2010.

References

Augustine, *De Mendacio*, http://www.augustinus.it/latino/menzogna/index.htm.

Cosmides, L. and Tooby, J. (2008), 'Can a General Deontic Logic Capture the Facts of Human Moral Reasoning? How the Mind Interprets Social Exchange Rules and Detects Cheaters', in Sinnott-Armstrong, W. (ed.), *Moral Psychology*, vol. 1. Cambridge Mass.: MIT Press, 53-119.

Ellis, R. D. (2010), 'On the Cusp', in Giordano, J. and Gordjin, B. (eds.), *Scientific and Philosophical Perspectives in Neuroethics*. Cambridge: Cambridge University Press, 66-94.

Greene, J. and Paxton, J. (2009), 'Patterns of Neural Activity Associated with Honest and Dishonest Moral Decisions', *Proceeding of the National Academy of Sciences*, 106 (30): 12506-12511.

Gurley, J. R. and Marcus, D. K. (2008), 'The Effects of Neuroimaging and Brain Injury on Insanity Defenses', *Behavioral Sciences and the Law*, 26: 85–97.

Hardcastle, V. G. and Stewart, C. M. (2009), 'fMRI: A Modern Cerebrascope? The Case of Pain', in Bickle, J. (ed.), *The Oxford Handbook of Philosophy and Neuroscience*, Oxford: Oxford University Press, 179-199.

Holderied-Milis, V. (2010), 'Online Chat – a Hatchery of Lies? Toward an Ethical Analysis of Truth and Lying in Internet Chatrooms', *Ethical Perspectives*, 1: 95-118.

Kanwisher, N. (2009), 'The Use of fMRI in Lie Detection: What Has Been Shown and What Has Not', in Bizzi, E., Hyman, S., Raichle, M., Kanwisher, N., Phelps, E., Morse, S. et al. (eds.), *Using Imaging to Identify Deceit*. Cambridge Mass.: American Academy of Arts and Sciences, 7-13.

Kim, J. (1998), *Mind in a Physical World*, London: MIT Press.

Madrigal, A. (2009.1), 'MRI Lie Detection to Get First Day in Court', *Wired Science*, 16 March, wired.com, 1-11.

Madrigal, A. (2009.2), 'Brain Scan Evidence Rejected by Brooklyn Court', *Wired Science*, 18 September, wired.com, 1-7.

Madrigal, A. (2010), 'Scanning Dead Salmon in fMRI Machine Highlights Risk of Red Herrings', *Wired Science*, 5 May, wired.com, 1-3.

McCabe, D. P., Castel, A. D. and Rhodes, M. G. (2011), 'The influence of fMRI Lie Detection Evidence on Juror Decision-Making', *Behavioral Sciences and the Law*, 29: 566-577.

Morse, S. (2006), 'Moral and Legal Responsibility and the New Neuroscience', in Illes, J. (ed.) *Neuroethics*. Oxford: Oxford University Press, 33-50.

Phelps, E. (2009), 'Lying Outside the Laboratory: The Impact of Imagery and Emotion on the Neural Circuitry of Lie Detection', in Bizzi, E. et al. (eds.), *Using Imaging to Identify Deceit*. American Academy of Arts and Sciences, 14-22.

Proust, M. (1999), *Albertine disparue*, in *A la recherche du temps perdu*, Paris: Gallimard-Quarto.

Raichle, M. (2009), 'An Introduction to Functional Brain Imaging in the Context of Lie Detection', in Bizzi, E. et al. (eds.), *Using Imaging to Identify Deceit*. American Academy of Arts and Sciences, 3-6.

Racine, E. (2010), *Pragmatic Neuroethics*, Cambridge Mass.: MIT Press.

Racine, E., Bar-Ilan, O. and Illes, J. (2005), 'fMRI in the Public Eye', *Nature Review Neuroscience*, 6 (2): 159-164.

Racine, E., Bell, E. and Illes, J. (2010), 'Can We Read Minds?', in Giordano, J. and Gordjin, B. (eds.), *Scientific and Philosophical Perspectives in Neuroethics*. Cambridge: Cambridge University Press, 244-270.

Rakoff, J. (2009), 'Lie Detection in the Courts: the Vain Search for the Magic Bullet', in Bizzi, E. et al. (eds.), *Using Imaging to Identify Deceit*. American Academy of Arts and Sciences, 40-45.

Roskies, A. L. (2007), 'Are Neuroimages Like Photographs of the Brain?', *Philosophy of Science*, 74: 860-872.

Saenz, A. (2010), 'Another Attempt to Use fMRI Lie Detector in US Courts Fails in Brooklyn', *Singularity Hub*, 6 May.

Sinnott-Armstrong, W., Roskies, A., Brown, T. and Murphy, E., (2008), 'Brain Images as Legal Evidence', *Episteme*, 359-373.

Sip, K. E., Roepstorff, A., McGregor, W. and Frith, C. D. (2007), 'Detecting Deception: The Scope and Limits', *Trends in Cognitive Neuroscience*, 12 (2): 48-53.

VanMeter, J. (2010). 'Neuroimaging', in Giordano, J. and Gordjin, B. (eds.), *Scientific and Philosophical Perspectives in Neuroethics*. Cambridge: Cambridge University Press, 230-243.

Weisberg, D. S., Keil, F. C., Goldstein, J. Rawson, E. and Gray, J. R. (2008), 'The Seductive Allure of Neuroscience Explanations', *Journal of Cognitive Neuroscience*, 20 (3): 470-477.

Wolpe, P. R., Foster, K. R. and Langleben, D. (2005), 'Emerging Neurotechnologies for Lie-Detection: Promises and Perils', *The American Journal f Bioethics*, 5 (2): 39-49.

A Measurement-Theoretic Foundation of Threshold Utility Maximiser's Preference Logic

Satoru Suzuki

Komazawa University, Japan

Abstract

There are at least two types of intransitivity of indifferences: multidimensional intransitivity and unidimensional intransitivity. We would face the Sorites Paradox if unidimensional indifferences were transitive. The Unidimensional Intransitivity Problem is as follows: what kind of preference logic can formalise inferences in which unidimensional indifferences are intransitive? In this paper, we explain the intransitivity of indifferences in terms of threshold utility maximisation. The aim of this paper is to propose a new version of complete and decidable preference logic—threshold utility maximiser's preference logic (TUMPL)— which can solve the Unidimensional Intransitivity Problem by means of measurement theory. The truth definition of TUMPL can guarantee that TUMPL is based on threshold utility maximisation. This truth definition can furnish a semantic solution to the Unidimensional Intransitivity Problem. A corollary of the Scott-Suppes theorem can relate threshold utility maximisation to semiorders, which enables us to propose the proof system of TUMPL on the basis of semiorders. This proof system can furnish a syntactic solution to the Unidimensional Intransitivity Problem.

Keywords: bounded rationality, intransitive indifference, measurement theory, semiorder, threshold utility maximisation

1. Introduction

The notion of preference plays an important role in many disciplines, including philosophy and economics. Hansson and Grüne-Yanoff (2006) conduct a comprehensive survey of preference in general. Some notable recent developments in ethics make substantial use of preference logic. For a comprehensive survey of preference logic, see Hansson (2001). In computer science, preference logic has become an indispensable device. The founder of preference logic is the founding father of logic itself, Aristotle. Book III of Aristotle's *Topics* can be regarded as the first treatment of this subject. From the 1950s to the 1960s, the study of preference logic flourished in Scandinavia—particularly by Halldén (1957) and Von Wright (1963)—and in the USA—particularly by Martin (1963) and Chisholm (1966). In recent years, using Boutilier's idea (Boutilier, 1994) that preferences between propositions can be defined in terms of two sorts of modalities, one of which is a universal modality, Van Benthem et al. (2005) reduce preference logic to modal logic. Van Benthem and Liu (2007) provide a logic of preference change, and Van Benthem et al. (2009) propose a logic for ceteris paribus preferences. In Suzuki (2009a), we propose a new version of sound and complete dynamic epistemic preference logic (DEPL).

The principle of Hansson (1968), Halldén (1957), and others that *indifference* is *transitive* has been criticised by many scholars. The economist Armstrong is one of the first to argue that indifference is not always transitive (Armstrong, 1939). Fishburn (1970, p.207) indicates six typical theories in which intransitive indifferences can appear:

1. basic preference theory,
2. consumer preference theory,
3. additive utility theory,
4. qualitative probability theory,

5. expected utility theory, and
6. social choice theory.

In both consumer preference theory and additive utility theory, the intransitivity of indifferences results from the *multidimensionality* of preferences and indifferences. In basic preference theory, which is *unidimensional*, on the other hand, the intransitivity of indifferences results from the fact that we cannot generally discriminate between very close quantities, or the fact that we are generally indifferent to the results of very fine discriminations. Fishburn (1970, p.208) gives the following example of the former:

Example 1 (Intransitivity of Multidimensional Indifferences) *If* $(x_1, \cdots, x_n)$ *and* $(y_1, \cdots, y_n)$ *differ only on one dimension (e.g.* $x_j = y_j$ *for all* $j > 1$ *and* $x_1 \neq y_1$*), even a small difference on this dimension may give rise to strict preference. But approximately offsetting differences on several dimensions may give rise to indifference areas that lead to intransitive indifference. We show this with a two-dimensional example suggested by the work of Armstrong and May. You are going to buy a car. You have no definite preference between (Ford, at* \$*2,600) and (Chevrolet, at* \$*2,700), and also have no definite preference between (Ford, at* \$*2,600) and (Chevrolet, at* \$*2,705). However, you prefer (Chevrolet, at* \$*2,700) to (Chevrolet, at* \$*2,705).*

On the other hand, Luce (1956, p.179) gives the following example of the latter:

Example 2 (Intransitivity of Unidimensional Indifferences) *Find a subject who prefers a cup of coffee with one cube of sugar to one with five cubes .... Now prepare 401 cups of coffee with* $(1 + \frac{i}{100})x$ *grams of sugar, for any* $i = 0, 1, \ldots, 400$*, where* x *is the weight of one cube of sugar. It is evident that he will be indifferent between cup* i *and cup* $i + 1$*, for any* i*, but by choice he is not indifferent between* $i = 0$ *and* $i = 400$*.*

This example has a lot in common with the Sorites Paradox. For a comprehensive survey of topics of vagueness, see Keefe (2000). The following argument is an ancient example of the Sorites Paradox:

Example 3 (Sorites Paradox) *1,000,000 grains of sand make a heap.*
If 1,000,000 grains of sand make a heap, then 999,999 grains of sand do.
If 999,999 grains of sand make a heap, then 999,998 grains do.
⋮
If 2 grains of sand make a heap, then 1 grain does.
1 grain of sand makes a heap.

Both Examples 2 and 3 show situations in which we would face a paradox if unidimensional indifferences (similarities) were *transitive*. One explanation for the intransitivity of unidimensional indifferences (similarities) is that this intransitivity results from the following:

1. the fact that we cannot generally discriminate between very close quantities, or
2. the fact that we are indifferent to the results of very fine discriminations.

Both Examples 2 and 3 illustrate the former. The following example (Ackerman, 1994, p.135) illustrates the latter.

Example 4 (Indifference as to Discrimination) *It is entirely plausible to suppose that an instructor would be indifferent to having the number of students in his seminar be 6 vs. 7, 7 vs. 8, etc., without being indifferent to having it be 6 vs. 15; he might consider 15 students too many for a seminar. But he can certainly discriminate between 6 and 7 students or 7 and 8, etc.*

A considerable number of studies have been carried out on preference logic. However, little attention has been paid to *preference logic* for intransitive indifferences, though numerous attempts have been made to study intransitive indifferences *themselves*. For further details on intransitive indifferences, see, for example, Fishburn (1970). Huber (1974, 1979) proposes preference logics for the intransitivity of multidimensional indifferences. However, completeness theorems and other important metatheorems remain to be proved for these logics. In this paper, we would like to focus on the following problem:

Problem 1 (Unidimensional Intransitivity Problem) *What kind of preference logic can formalise inferences in which unidimensional indifferences are intransitive?*

We call this the *Unidimensional Intransitivity Problem*. The aim of this paper is to propose a new version of complete and decidable preference logic—*threshold utility maximiser's preference logic* (TUMPL)—which can solve the Unidimensional Intransitivity Problem in terms of *measurement theory*. For a comprehensive survey of measurement theory, see Roberts (1979).

The structure of this paper is as follows. In Section 2, we make some observations on bounded rationality, intransitive indifferences, and the accessibility relation. In Section 3, we define the language $\mathcal{L}_{\mathsf{TUMPL}}$ of TUMPL. In Section 4, we define a structured Kripke model $\mathfrak{M}$ for TUMPL, and provide TUMPL with a truth definition. In Section 5, we provide TUMPL with a proof system. In Section 6, we sketch the proof of the soundness, completeness, and decidability of TUMPL.

2. Bounded Rationality: Intransitive Indifferences and Accessibility Relation

2.1 Intransitive Indifferences and Irrationality

Most proof systems of preference logic include a transitivity axiom. This is motivated not only by mathematical convenience, but primarily by the fact that transitivity of preferences is a compelling requirement of preferences. Tversky (1969, p.455) makes the following remark on the relation between the transitivity of preferences and the Money Pump Argument:

> Transitivity, however, is one of the basic and the most compelling principles of rational behavior. For if one violates transitivity, it is a well-known conclusion that he is acting, in effect, as a "money-pump."

We can trace the origin of the Money Pump Argument back to Davidson et al. (1955). Tversky (1969, pp.455–456) goes on to say on a form of the Money Pump Argument:

> Suppose an individual prefers y to x, z to y, and x to z. It is reasonable to assume that he is willing to pay a sum of money to replace x by y. Similarly, he should be willing to pay some amount of money to replace y by z and still a third amount to replace z by x. Thus, he ends up with the alternative he started with but with less money.

Lehrer and Wagner (1985, pp.249–250) give an example of the Money Pump Argument against the intransitivity of indifferences:

> Let us consider the case of a buyer of wine. This individual, after extensive tasting, finds that he is equally attracted to wines A and B, and to wines B and C, yet prefers A to C. … It turns out that wines A and B are available on Monday and wine C on Tuesday. As the buyer is indifferent between A and B, the merchant provisionally chooses B for him, and since the buyer is indifferent between B and C, the merchant makes for him the final choice of C. Thus, although A was available and preferred to C, the buyer receives C, in full compliance with his instructions.

Does the intransitivity of indifferences always lead to irrationality? This argument is considered to assume 'diachronic independence (additivity)'. Diachronic independence is as follows:

> $\cdots$ that the arrangements are value-wise independent, that if the agent knew of the arrangements he had already accepted, this would not affect the value he set on the arrangement just offered him (Schick, 1986, p.117).

We agree with Schick's statement that:

> Again, the additivity/independence assumption cannot be taken for granted. Indeed, in typical cases it is false, and for the obvious reasons: the gradual depletion of the agent's funds, his awareness of being exploited, and the like (Schick, 1986, p.117).

Therefore, because the independence assumption is not valid, Lehrer and Wagner's argument does not establish that the intransitivity of indifferences is irrational. Then, in what sense can we say that the intransitivity of indifferences is rational?

2.2 Intransitive Indifferences and Bounded Rationality

The standard model of economics is based on *global rationality*, which requires an *optimising behaviour*. *Utility maximisation* is a typical example of an optimising behaviour. When a set $\mathbf{A}$ of objects and a utility function $u : \mathbf{A} \to \mathbb{R}$ are given, utility maximisation is formulated as follows:

> For any $x, y \in \mathbf{A}$, the subject weakly prefers x to y iff $u(x) \geq u(y)$.

However, according to Simon (1982), cognitive and information-processing constraints on the capabilities of agents, together with the complexity of their environment, render an optimising behaviour an unattainable ideal. He dismisses the idea that agents should exhibit global rationality and suggests that they in fact exhibit *bounded rationality*, which allows a *satisficing behaviour*. On the relation between the Sorites Paradox and bounded rationality, see Van Rooij (2011). As stated in the Introduction, one explanation for Examples 2 and 3 is that the intransitivity of indifferences results from (1) the fact that we cannot generally discriminate between very close quantities, or (2) the fact that we are indifferent to the results of very fine discriminations. The psychophysicist Fechner explains this inability (indifference) through the concept of *threshold* of discrimination, that is, *just noticeable difference* (JND) (Fechner, 1860). In Suzuki (2011a, b), we propose two versions of logic for vague predicates, each of whose models is based on JNDs. Given the measure function f that the experimenter assigns to the subject and the object a, its JND δ is the *lowest intensity increment* such that $f(a) + \delta$ is recognised as higher than $f(a)$ by the subject. The notion of JND is closely related to *threshold utility maximisation*. When a set $\mathbf{A}$ of objects, a utility function $u : \mathbf{A} \to \mathbb{R}$, and a positive threshold δ are given, threshold utility maximisation is formulated as follows:

- For any $x, y \in \mathbf{A}$, the subject strictly prefers x to y iff $u(x) > u(y) + \delta$.
- For any $x, y \in \mathbf{A}$, the subject is indifferent between x and y iff $u(x) \not> u(y) + \delta$ and $u(y) \not> u(x) + \delta$.

In this paper, we explain the intransitivity of indifferences in terms of threshold utility maximisation. Because threshold utility maximisation is based on the limited ability of the subject, it is considered to be a typical example of a satisficing behaviour. In this sense, we can say that the intransitivity of indifferences is boundedly rational.

2.3 Bounded Rationality and Accessibility Relation

The language of TUMPL includes a necessity operator $\Box$. The aim of introducing $\Box$ is as follows. Let $\mathcal{W}$ be a non-empty set of possible worlds. Suppose that the subject is boundedly rational. Then, in each $w \in \mathcal{W}$, he may have different universes, $\mathcal{W}_w$, relative to w for considering preferences. For he might not necessarily be able to take any $w \in \mathcal{W}$ into consideration because of some limitation in his ability. To make this point explicit, we introduce an accessibility relation R on $\mathcal{W}$. By means of R, we can define $\mathcal{W}_w$ as $\{w' \in \mathcal{W} : R(w, w')\}$. $R(w, w')$ is interpreted to mean that in w, he can imagine $w' \in \mathcal{W}$ as an alternative to w. We need $\Box$ to verbalise the behaviour of R in TUMPL. However, it must be noted that we introduce R for the purpose not of dealing with cases like Examples 2 and 3, but of representing bounded rationality of the subject.

3. Language of Threshold Utility Maximiser's Preference Logic

We define the language $\mathcal{L}_{\mathsf{TUMPL}}$ of TUMPL.

Definition 1 (Language) *Let $\mathcal{S}$ denote a set of sentential variables, $\Box$ a unary sentential operator, and* **SPR** *a binary sentential operator. The language $\mathcal{L}_{\mathsf{TUMPL}}$ of* TUMPL *is given by the following rule:*

$$\varphi ::= s \mid \top \mid \neg\varphi \mid (\varphi \& \varphi) \mid \Box\varphi,$$
$$\psi ::= \varphi \mid \neg\psi \mid (\psi \& \psi) \mid \Box\psi \mid \mathbf{SPR}(\varphi, \varphi)$$

such that $s \in \mathcal{S}$.

- *$\bot, \vee, \to, \leftrightarrow$ and $\Diamond$ are introduced by the standard definitions.*
- *$\Box\varphi$ is interpreted to mean that necessarily φ.*

We define an indifference relation symbol **IND** *and a weak preference relation symbol* **WPR** *as follows:*

- $\mathbf{IND}(\varphi, \psi) := \neg\mathbf{SPR}(\varphi, \psi) \& \neg\mathbf{SPR}(\psi, \varphi)$.
- $\mathbf{WPR}(\varphi, \psi) := \mathbf{SPR}(\varphi, \psi) \vee \mathbf{IND}(\varphi, \psi)$.

The set of all well-formed formulae of $\mathcal{L}_{\mathsf{TUMPL}}$ will be denoted by $\Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$.

4. Semantics of Threshold Utility Maximiser's Preference Logic

In this section, we prove that the truth definition of TUMPL can guarantee that TUMPL is based on *threshold utility maximisation*. This truth definition can furnish a semantic solution to the Unidimensional Intransitivity Problem.

4.1 Model

We define a structured Kripke model $\mathfrak{M}$ for TUMPL.

Definition 2 (Model) *$\mathfrak{M}$ is a quadruple $(\mathcal{W}, R, V, \rho)$ in which*

- *$\mathcal{W}$ is a non-empty set of possible worlds,*
- *R is a binary accessibility relation on $\mathcal{W}$,*
- *V is a truth assignment to each $s \in \mathbf{S}$ for each $w \in \mathcal{W}$, and*
- *ρ is a threshold utility space assignment that assigns to each $w \in \mathcal{W}$ a threshold utility space $(\mathcal{W}_w, \mathcal{F}_w, U_w, \delta_w)$ in which*
  - $\mathcal{W}_w := \{w' \in \mathcal{W} : R(w, w')\}$,
  - *$\mathcal{F}_w$ is a Boolean algebra of subsets of $\mathcal{W}_w$ with $\emptyset$ as zero element and $\mathcal{W}_w$ as unit element,*
  - *U_w is a utility function from $\mathcal{F}_w$ to $\mathbb{R}$ that the experimenter assigns to the subject, and*
  - *δ_w is a positive threshold (a JND) relative to w such that $U_w(\alpha) + \delta_w$ is recognised as higher than $U_w(\alpha)$ by the subject, for any $\alpha \in \mathcal{F}_w$.*

Remark 1 *$R(w, w')$ above is interpreted to mean that, in $w \in \mathcal{W}$, the subject can imagine $w' \in \mathcal{W}$ as an alternative to w. $\mathcal{W}_w$ is interpreted as the universe relative to w in which the subject considers the utility of propositions, and $\mathcal{F}_w$ is interpreted as the set of propositions that, in $w \in \mathcal{W}$, the subject takes into consideration.*

4.2 Truth

We provide TUMPL with the following truth definition relative to $\mathfrak{M}$:

Definition 3 (Truth) *The notion of $\varphi \in \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$ being true at $w \in W$ in $\mathfrak{M}$, in symbols $(\mathfrak{M}, w) \models_{\mathsf{TUMPL}} \varphi$ is inductively defined as follows:*

- $(\mathfrak{M}, w) \models_{\mathsf{TUMPL}} s$ *iff* $V(w)(s) = \textbf{true}$,

- $(\mathfrak{M}, w) \models_{\mathsf{TUMPL}} \top$,

- $(\mathfrak{M}, w) \models_{\mathsf{TUMPL}} \varphi \& \psi$ *iff* $(\mathfrak{M}, w) \models_{\mathsf{TUMPL}} \varphi$ *and* $(\mathfrak{M}, w) \models_{\mathsf{TUMPL}} \psi$,

- $(\mathfrak{M}, w) \models_{\mathsf{TUMPL}} \neg\varphi$ *iff* $(\mathfrak{M}, w) \not\models_{\mathsf{TUMPL}} \varphi$,

- $(\mathfrak{M}, w) \models_{\mathsf{TUMPL}} \Box\varphi$ *iff, for any w' such that $R(w, w')$,* $(\mathfrak{M}, w') \models_{\mathsf{TUMPL}} \varphi$,

and

- $(\mathfrak{M}, w) \models_{\mathsf{TUMPL}} \mathbf{SPR}(\varphi, \psi)$ *iff* $U_w([\![\varphi]\!]^{\mathfrak{M}}_w) > U_w([\![\psi]\!]^{\mathfrak{M}}_w) + \delta_w$

in which $[\![\varphi]\!]^{\mathfrak{M}}_w := \{w' \in \mathcal{W} : R(w, w')$ and $(\mathfrak{M}, w') \models_{\mathsf{TUMPL}} \varphi\}$. If $(\mathfrak{M}, w) \models_{\mathsf{TUMPL}} \varphi$ for all $w \in \mathcal{W}$, we write $\mathfrak{M} \models_{\mathsf{TUMPL}} \varphi$ and say that φ is valid in $\mathfrak{M}$. If φ is valid in all structured Kripke models for TUMPL, *we write $\models_{\mathsf{TUMPL}} \varphi$ and say that φ is valid.*

Remark 2 *This truth definition can furnish a semantic solution to the Unidimensional Intransitivity Problem.*

The truth condition of **IND** follows directly from Definition 1 and Definition 3.

Corollary 1

$$
\begin{aligned}
&(\mathfrak{M}, w) \models_{\mathsf{TUMPL}} \mathbf{IND}(\varphi, \psi) \\
&\textit{iff} \quad U_w([\![\varphi]\!]^{\mathfrak{M}}_w) \not> U_w([\![\psi]\!]^{\mathfrak{M}}_w) + \delta_w \textit{ and} \\
&\qquad\; U_w([\![\psi]\!]^{\mathfrak{M}}_w) \not> U_w([\![\varphi]\!]^{\mathfrak{M}}_w) + \delta_w.
\end{aligned}
$$

4.3 Counter-Model

We can provide a *counter-model* that falsifies the transitivity of indifferences. We now return to Example 2. Assume that $\mathfrak{T} := (\mathcal{W}, R, V, \rho)$ is given in which

- $\mathcal{W} := \{w_0, \ldots, w_{400}\}$ such that w_i is a possible world in which the subject tries a cup of coffee with $(1 + \frac{i}{100})x$ grams of sugar, for any i $(0 \leq i \leq 400)$,

- R is a binary accessibility relation on $\mathcal{W}$, where $R(w_i, w_j)$ is interpreted to mean that, in $w_i \in \mathcal{W}$, the subject can imagine $w_j \in \mathcal{W}$ as an alternative to w_i when he tries a cup of coffee,

- V is a truth assignment to each $s \in \mathbf{S}$ for each $w_i \in \mathcal{W}$, and

- ρ is a threshold utility space assignment that assigns to each $w_i \in \mathcal{W}$ a threshold utility space $(\mathcal{W}_{w_i}, \mathcal{F}_{w_i}, U_{w_i}, \delta_{w_i})$ in which
  - $\mathcal{W}_{w_i} := \{w_j \in \mathcal{W} : R(w_i, w_j)\}$, where $\mathcal{W}_{w_i}$ is interpreted as the universe relative to w_i in which the subject considers the utility of propositions when he tries a cup of coffee,
  - $\mathcal{F}_{w_i}$ is a Boolean algebra of subsets of $\mathcal{W}_{w_i}$ with $\emptyset$ as zero element and $\mathcal{W}_{w_i}$ as unit element, where $\mathcal{F}_{w_i}$ is interpreted as the set of propositions that, in $w_i \in \mathcal{W}$, the subject takes into consideration when he tries a cup of coffee,
  - U_{w_i} is a utility function from $\mathcal{F}_{w_i}$ to $\mathbb{R}$ that the experimenter assigns to the subject when the latter tries a cup of coffee,
  - δ_{w_i} is a positive threshold (a JND) relative to w_i such that $U_{w_i}(\alpha) + \delta_{w_i}$ is recognised as higher than $U_{w_i}(\alpha)$ by the subject, for any $\alpha \in \mathcal{F}_{w_i}$, when he tries a cup of coffee, and
  - U_{w_i} and δ_{w_i} satisfy the following conditions:

$$
\begin{cases}
U_{w_i}(\{w_j\}) \not> U_{w_i}(\{w_{j+1}\}) + \delta_{w_i}, \\
U_{w_i}(\{w_{j+1}\}) \not> U_{w_i}(\{w_j\}) + \delta_{w_i},
\end{cases}
$$
$$\text{for any } j \ (0 \leq j \leq 400),$$

$$U_{w_i}(\{w_0\}) > U_{w_i}(\{w_{400}\}) + \delta_{w_i}.$$

Because, in any $w_i \in \mathcal{W}$, the subject can imagine any $w_j \in \mathcal{W}$ as an alternative to w_i when he tries a cup of coffee, we have

$$\text{For any } w_i, w_j \in \mathcal{W},\ R(w_i, w_j).$$

So, by the definition of W_{w_i}, we have

$$\mathcal{W}_{w_0} = \cdots = \mathcal{W}_{w_{400}} = \mathcal{W}.$$

Because all the relative universes are the same, it is plausible to suppose that all the relative sets of propositions that the subject takes into consideration when he tries a cup of coffee are the same, that is,

$$\mathcal{F}_{w_0} = \cdots = \mathcal{F}_{w_{400}}.$$

Let φ_i denote the sentence 'The subject tries a cup of coffee with $(1 + \frac{i}{100})x$ grams of sugar', for any i $(0 \leq i \leq 400)$. Then, we have, for any i $(0 \leq i \leq 400)$,

$$[\![\varphi_j]\!]^{\mathfrak{M}}_{w_i} = \{w_j\}, \text{ for any } j \ (0 \leq j \leq 400).$$

So, we have, for any i $(0 \leq i \leq 400)$,

$$
(a) \quad
\begin{cases}
U_{w_i}([\![\varphi_j]\!]^{\mathfrak{M}}_{w_i}) \not> U_{w_i}([\![\varphi_{j+1}]\!]^{\mathfrak{M}}_{w_i}) + \delta_{w_i}, \\
U_{w_i}([\![\varphi_{j+1}]\!]^{\mathfrak{M}}_{w_i}) \not> U_{w_i}([\![\varphi_j]\!]^{\mathfrak{M}}_{w_i}) + \delta_{w_i},
\end{cases}
$$
$$\text{for any } j \ (0 \leq j \leq 400),$$

$$(b) \quad U_{w_i}([\![\varphi_0]\!]^{\mathfrak{M}}_{w_i}) > U_{w_i}([\![\varphi_{400}]\!]^{\mathfrak{M}}_{w_i}) + \delta_{w_i}.$$

Because (a) does not imply (b), we have, for any i ($0 \leq i \leq 400$),

$$(\mathfrak{T}, w_i) \not\models_{\mathsf{TUMPL}} (\mathbf{IND}(\varphi_0, \varphi_1) \& \cdots \& \mathbf{IND}(\varphi_{399}, \varphi_{400})) \\ \to \mathbf{IND}(\varphi_0, \varphi_{400}).$$

Therefore, we obtain the following proposition.

Proposition 1 (Intransitivity of Indifferences)

$$\not\models_{\mathsf{TUMPL}} (\mathbf{IND}(\varphi_0, \varphi_1) \& \cdots \& \mathbf{IND}(\varphi_{399}, \varphi_{400})) \\ \to \mathbf{IND}(\varphi_0, \varphi_{400}).$$

5. Syntax of Threshold Utility Maximiser's Preference Logic

In this section, we prove that a corollary of the *Scott-Suppes theorem* can relate threshold utility maximisation to semiorders, which enables us to propose the proof system of TUMPL on the basis of *semiorders*. This proof system can furnish a syntactic solution to the Unidimensional Intransitivity Problem.

5.1 From Semantics to Syntax: Semiorder, Weak Order, and Representation Theorems

Luce (1956) introduces the concept of a *semiorder* that can provide a *qualitative* counterpart of a JND that is quantitative. Scott and Suppes (1958, p.117) define a semiorder as follows:

Definition 4 (Semiorder) *A binary relation $\succ$ on $\mathbf{A}$ is called a semiorder if, for any $w, x, y, z \in \mathbf{A}$, the following conditions are satisfied:*

1. *$x \not\succ x$ (Irreflexivity),*
2. *If $w \succ x$ and $y \succ z$, then $w \succ z$ or $y \succ x$ (Strong Intervality), and*
3. *If $w \succ x$ and $x \succ y$, then $w \succ z$ or $z \succ y$ (Semitransitivity).*

There are two main problems with *measurement theory*:

1. the *representation problem*—justifying the assignment of numbers to objects,
2. the *uniqueness problem*—specifying the transformation up to which this assignment is unique.

A solution to the former can be furnished by a *representation theorem*, which establishes that the specified conditions on a qualitative relational system are (necessary and) sufficient for the assignment of numbers to objects which represents (or preserves) all the relations in the system.

Cantor (1895) proves the representation theorem that can relate *utility maximisation* to *weak orders*.

Theorem 1 (Cantor, 1895) *Suppose $\mathbf{A}$ is a countable set and $\succeq$ is a binary relation on $\mathbf{A}$. Then, $\succeq$ is a weak order (transitive and connected) iff there is a function $u : \mathbf{A} \to \mathbb{R}$ such that for any $x, y \in \mathbf{A}$,*

$$x \succeq y \text{ iff } u(x) \geq u(y).$$

Scott and Suppes (1958) prove a representation theorem for semiorders when $\mathbf{A}$ is *finite*.

Theorem 2 (Scott and Suppes, 1958) *Suppose that $\succ$ is a binary relation on a finite set $\mathbf{A}$ and δ is a positive number. Then $\succ$ is a semiorder iff there is a function $u : \mathbf{A} \to \mathbb{R}$ such that for any $x, y \in \mathbf{A}$,*

$$x \succ y \text{ iff } u(x) > u(y) + \delta.$$

Remark 3 *Scott (1964) simplifies the Scott-Suppes theorem in terms of the solvability of the finite system of linear inequalities.*

Since A is an *arbitrary* finite set, the next corollary follows directly from Theorem 2.

Corollary 2 (Representation on Finite Boolean Algebra) *Suppose that $\mathcal{W}$ is a finite set of possible worlds, $\mathcal{F}$ is a finite Boolean algebra of subsets of $\mathcal{W}$, $\succ$ is a binary relation on $\mathcal{F}$, and δ is a positive number. Then, $\succ$ is a semiorder iff there is a function $U : \mathcal{F} \to \mathbb{R}$ such that for any $\alpha, \beta \in \mathcal{F}$,*

$$\alpha \succ \beta \text{ iff } U(\alpha) > U(\beta) + \delta.$$

5.2 Proof System

Corollary 2 can relate *threshold utility maximisation* to *semiorders*, which enables us to propose the following proof system of TUMPL on the basis of semiorders.

Definition 5 (Proof System) *The proof system of* TUMPL *consists of the following:*

1. *all tautologies of classical sentential logic,*
2. $\Box(\varphi_1 \to \varphi_2) \to (\Box\varphi_1 \to \Box\varphi_2)$ (K),
3. $\Box(\varphi_1 \leftrightarrow \varphi_2) \& \Box(\psi_1 \leftrightarrow \psi_2) \to (\mathbf{SPR}(\varphi_1, \psi_1) \leftrightarrow \mathbf{SPR}(\varphi_2, \psi_2))$ *(Replacement of Necessary Equivalents),*
4. $\neg\mathbf{SPR}(\varphi, \varphi)$ *(Syntactic Counterpart of Irreflexivity),*
5. $(\mathbf{SPR}(\varphi_1, \varphi_2) \& \mathbf{SPR}(\varphi_3, \varphi_4)) \to (\mathbf{SPR}(\varphi_1, \varphi_4) \vee \mathbf{SPR}(\varphi_3, \varphi_2))$ *(Syntactic Counterpart of Strong Intervality),*
6. $(\mathbf{SPR}(\varphi_1, \varphi_2) \& \mathbf{SPR}(\varphi_2, \varphi_3)) \to (\mathbf{SPR}(\varphi_1, \varphi_4) \vee \mathbf{SPR}(\varphi_4, \varphi_3))$ *(Syntactic Counterpart of Semitransitivity),*

7. *Modus Ponens, and*

8. *Necessitation.*

A proof of $\varphi \in \Phi_{\mathsf{TUMPL}}$ is a finite sequence of $\mathcal{L}_{\mathsf{TUMPL}}$-formulae having φ as the last formula such that either each formula is an instance of an axiom or it can be obtained from formulae that appear earlier in the sequence by applying an inference rule. If there is a proof of φ, we write $\vdash_{\mathsf{TUMPL}} \varphi$.

Remark 4 *This proof system can furnish a syntactic solution to the Unidimensional Intransitivity Problem.*

6. Metalogic of Threshold Utility Maximiser's Preference Logic

We prove the metatheorems of TUMPL. It is easy to prove the soundness of TUMPL.

Theorem 3 (Soundness) *For any $\varphi \in \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$, if $\vdash_{\mathsf{TUMPL}} \varphi$, then $\models_{\mathsf{TUMPL}} \varphi$.*

We now turn to the task of proving the completeness of TUMPL. We prove it by using the ideas of Segerberg (1968, 1971) and modifying *filtration* in such a way that completeness can be established by Corollary 2. We cannot go into detail because of limited space, but the outline of the proof is as follows. We begin by defining some new concepts.

Definition 6 (Stuffedness) *Suppose that Θ is a set of formulae such that Θ is closed under subformulae. Let*

$$\Delta := \{\varphi : \textit{for some } \psi,\ \mathbf{SPR}(\varphi,\psi) \in \Theta \textit{ or } \mathbf{SPR}(\psi,\varphi) \in \Theta\},$$

and let Δ' be the closure of Δ under Boolean compounds. If Θ also satisfies the condition that $\mathbf{SPR}(\varphi,\psi) \in \Theta$, for any $\varphi,\psi \in \Delta'$, we say that Θ is stuffed.

Definition 7 (Value Formula) *The formulae in Δ' are called the value formulae of Θ.*

Remark 5 *There is no occurrence of* **SPR** *in value formulae.*

Definition 8 (Base) *We say that $\Psi_0 \subseteq \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$ is a base (with respect to* TUMPL*) for $\Psi \subseteq \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$ if for any $\varphi \in \Psi$ there is some $\varphi_0 \in \Psi_0$ such that $\vdash_{\mathsf{TUMPL}} \varphi \leftrightarrow \varphi_0$.*

Definition 9 (Logical Finiteness) *We say that Ψ is logically finite (with respect to* TUMPL*) if there is a finite base for Ψ.*

Lemma 1 (Logical Finiteness) *If $\Psi \subseteq \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$ is a finite set closed under subformulae, and if Θ is the smallest stuffed superset of Ψ, then Θ is logically finite.*

Definition 10 (Maximal Consistency) *A finite set $\{\varphi_1, \ldots, \varphi_n\} \subseteq \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$ is* TUMPL*-consistent iff $\nvdash_{\mathsf{TUMPL}} \neg(\varphi_1 \& \ldots \& \varphi_n)$. An infinite set of formulae is* TUMPL*-consistent iff all of its finite subsets are* TUMPL*-consistent. $\Gamma \subseteq \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$ is a* TUMPL*-maximal consistent set iff it is* TUMPL*-consistent and for any $\varphi \notin \Gamma$, $\Gamma \cup \{\varphi\}$ is* TUMPL*-inconsistent.*

Definition 11 (Canonical Model for Alethic-Modal Part) *We define $\mathfrak{U}^C := (\mathcal{X}^C, R^C, V^C)$ as a canonical model for the alethic-modal part of* TUMPL *in which*

- *$\mathcal{X}^C := \{\Gamma \subseteq \Phi_{\mathcal{L}_{\mathsf{TUMPL}}} : \Gamma$ is* TUMPL*-maximal consistent$\}$,*
- *for any $\Gamma, \Delta \in \mathcal{X}^C$, $R^C(\Gamma, \Delta)$ iff for any $\varphi \in \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$, if $\Box\varphi \in \Gamma$, then $\varphi \in \Delta$, and*
- *for any $\Gamma \in \mathcal{X}^C$,*

$$V^C(\Gamma)(s) := \begin{cases} \textbf{\textit{true}} & \textit{if } s \in \Gamma, \\ \textbf{\textit{false}} & \textit{otherwise.} \end{cases}$$

Definition 12 (Equivalence Class) *Let Θ be a stuffed set of formulae that are logically finite with respect to* TUMPL*. We define, for $\Gamma, \Delta \in \mathcal{X}^C$,*

$$\Gamma \equiv_\Theta \Delta \textit{ iff } \Gamma \cap \Theta = \Delta \cap \Theta.$$

Then, $\equiv_\Theta$ is an equivalence relation modulo Θ on $\mathcal{X}^C$. We write $[\Gamma]_\Theta$ for the equivalence class of Γ under $\equiv_\Theta$.

Definition 13 (Filtration) *We define $\mathfrak{U}^\Theta := (\mathcal{X}^\Theta, R^\Theta, V^\Theta)$ as a filtration of $\mathfrak{U}^C$ through Θ in which*

- *$\mathcal{X}^\Theta := \{[\Gamma]_\Theta : \Gamma \in \mathcal{X}^C\}$,*
- *R^Θ is a binary relation on $\mathcal{X}^\Theta$ such that*
    1. *if $R^C(\Gamma, \Delta)$, then $R^\Theta([\Gamma]_\Theta, [\Delta]_\Theta)$,*
    2. *if $R^\Theta([\Gamma]_\Theta, [\Delta]_\Theta)$ and $\Box\varphi \in \Gamma$, then $\varphi \in \Delta$, and*
- *V^Θ is a function such that for any $s \in \Theta$,*

$$V^\Theta([\Gamma]_\Theta)(s) = V^C(\Gamma)(s).$$

Thus, for any $\xi \in \mathcal{X}^\Theta$,

$$[\![\varphi]\!]_\xi^{\mathfrak{U}^\Theta} := \{\eta : R^\Theta(\xi, \eta) \text{ and } (\mathfrak{U}^\Theta, \eta) \models_{\mathsf{TUMPL}} \varphi\}$$

is well-defined for any φ that does not contain **SPR**.

Lemma 2 (Lindenbaum) *Every* TUMPL*-consistent set of formulae is a subset of a* TUMPL*-maximal consistent set of formulae.*

Lemma 3 (Partial Truth) *If $\varphi \in \Theta$ and φ does not contain* **SPR***, then for any $\Gamma \in \mathcal{X}^C$,*

$$(\mathfrak{U}^\Theta, [\Gamma]_\Theta) \models_{\mathsf{TUMPL}} \varphi \textit{ iff } \varphi \in \Gamma.$$

We wish to supplement $\mathfrak{U}^{\Theta}$ with a threshold utility space assignment ρ^{Θ} so as to obtain a structured Kripke model $\mathfrak{U}^{\Theta}_{\sharp}$ for which Truth Lemma holds for all formulae in Θ. Doing this contributes to solving the completeness problem of TUMPL.

Definition 14 ($\mathcal{F}^{\Theta}_{\xi}$) *For any $\xi \in \mathcal{X}^{\Theta}$, we define $\mathcal{F}^{\Theta}_{\xi}$ as the set of all $\alpha \subseteq \mathcal{X}^{\Theta}_{\xi} := \{\eta : R^{\Theta}(\xi,\eta)\}$ such that for some value formula $\varphi \in \Theta$, $\alpha = [\![\varphi]\!]^{\mathfrak{U}^{\Theta}}_{\xi}$.*

Lemma 4 (Boolean Algebra) *For any $\xi \in \mathcal{X}^{\Theta}$, $\mathcal{F}^{\Theta}_{\xi}$ is a Boolean algebra with $\emptyset$ as zero element and $\mathcal{X}^{\Theta}_{\xi}$ as unit element.*

Definition 15 ($\succ_{\xi}$) *For any $\xi \in \mathcal{X}^{\Theta}$, we define $\alpha \succ_{\xi} \beta$ to hold between elements $\alpha,\beta \in \mathcal{F}^{\Theta}_{\xi}$ iff there are value formulae $\varphi,\psi \in \Theta$ such that $\alpha = [\![\varphi]\!]^{\mathfrak{U}^{\Theta}}_{\xi}$, $\beta = [\![\psi]\!]^{\mathfrak{U}^{\Theta}}_{\xi}$ and* $\mathbf{SPR}(\varphi,\psi) \in \Gamma$ *for any $\Gamma \in \xi$.*

Lemma 5 ($\succ_{\xi}$ and SPR) *For any value formula $\varphi,\psi \in \Theta$ and any $\xi \in \mathcal{X}^{\Theta}$, $[\![\varphi]\!]^{\mathfrak{U}^{\Theta}}_{\xi} \succ_{\xi} [\![\psi]\!]^{\mathfrak{U}^{\Theta}}_{\xi}$ iff, for any $\Gamma \in \xi$,* $\mathbf{SPR}(\varphi,\psi) \in \Gamma$.

The next lemma follows from Lemma 5.

Lemma 6 (Conditions for Semiorders) *For any $\xi \in \mathcal{X}^{\Theta}$, $\succ_{\xi}$ on $\mathcal{F}^{\Theta}_{\xi}$ satisfies Irreflexivity, Strong Intervality, and Semitransitivity.*

Since we assumed that Θ is logically finite, $\mathcal{X}^{\Theta}$ is *finite*. Hence for any $\xi \in \mathcal{X}^{\Theta}$, $\mathcal{F}^{\Theta}_{\xi}$ is *finite*, so the next corollary follows from Corollary 2, Lemma 4, and Lemma 6.

Corollary 3 (Representation on $\mathcal{F}^{\Theta}_{\xi}$) *For any $\xi \in \mathcal{X}^{\Theta}$, there is a utility function $U_{\xi} : \mathcal{F}^{\Theta}_{\xi} \to \mathbb{R}$ such that for any $\alpha,\beta \in \mathcal{F}^{\Theta}_{\xi}$,*

$$\alpha \succ_{\xi} \beta \text{ iff } U_{\xi}(\alpha) > U_{\xi}(\beta) + \delta_{\xi}.$$

Definition 16 ($\mathfrak{U}^{\Theta}_{\sharp}$) *We define $\mathfrak{U}^{\Theta}_{\sharp}$ as $(\mathcal{X}^{\Theta}, R^{\Theta}, V^{\Theta}, \rho^{\Theta})$ in which ρ^{Θ} is a threshold utility space assignment that assigns $(\mathcal{X}^{\Theta}_{\xi}, \mathcal{F}^{\Theta}_{\xi}, U_{\xi}, \delta_{\xi})$ to each $\xi \in \mathcal{X}^{\Theta}$.*

Lemma 7 (Full Truth) *For any $\varphi \in \Theta$ and any $\Gamma \in \mathcal{X}^{C}$,*

$$(\mathfrak{U}^{\Theta}_{\sharp}, [\Gamma]_{\Theta}) \models_{\mathsf{TUMPL}} \varphi \text{ iff } \varphi \in \Gamma.$$

Remark 6 *This lemma is the announced improvement of Lemma 3.*

Theorem 4 (Completeness) *For any $\varphi \in \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$, if $\models_{\mathsf{TUMPL}} \varphi$, then $\vdash_{\mathsf{TUMPL}} \varphi$.*

Proof Suppose that $\nvdash_{\mathsf{TUMPL}} \varphi_0$. Then $\{\neg\varphi_0\}$ is a TUMPL-consistent set. By Lemma 2, $\{\neg\varphi_0\}$ is a subset of a TUMPL-maximal consistent set Γ. Evidently, $\varphi_0 \notin \Gamma$. Let Ψ be the set of subformulae of TUMPL which is finite and let Θ be the smallest stuffed extension of Ψ. By Lemma 1, Θ is logically finite with respect to TUMPL. If $\mathfrak{U}^{\Theta}_{\sharp}$ is constructed as above, it follows from Lemma 7 that $(\mathfrak{U}^{\Theta}_{\sharp}, [\Gamma]_{\Theta}) \not\models_{\mathsf{TUMPL}} \varphi_0$. Therefore, $\not\models_{\mathsf{TUMPL}} \varphi_0$. □

We can prove the decidability of TUMPL as follows.

Lemma 8 (Finite Model Property) TUMPL *has the finite model property that every non-theorem of* TUMPL *fails in a structured Kripke model for* TUMPL *with only a finite number of elements.*

Theorem 5 (Decidability) TUMPL *is decidable.*

Proof Suppose that φ is not provable in TUMPL. By Lemma 8, φ fails in a structured Kripke model $\mathfrak{U}^{\Theta}_{\sharp}$ for TUMPL with a finite number of elements. If we take a domain $\mathcal{X}^{\Theta}$ with, at most, that many elements, there are only a finite number of ways in which accessibility relations and truth assignments can be defined, and there are also only a finite number of ways to define the threshold utility space assignment ρ^{Θ}. Whether a defined relation, $\succ_{\xi}$, satisfies Irreflexivity, Strong Intervality, and Semitransitivity can be decided in a finite number of steps. Thus, we find, in at most a finite number of steps, a counter-model that falsifies the unprovable formula. In fact, we can compute an upper bound to the number of steps needed. Thus, TUMPL is decidable. □

7. Concluding Remarks

In this paper, we have proposed a new version of complete and decidable preference logic—threshold utility maximiser's preference logic (TUMPL)—which can solve the Unidimensional Intransitivity Problem.

This paper is only a part of a larger measurement-theoretic study. We are now trying to construct such logics as dynamic epistemic preference logic (Suzuki, 2009a), dyadic deontic logic (Suzuki, 2009b), a logic for goodness and badness (Suzuki, 2009c), vague predicate logic (Suzuki, 2011a, b), a logic of interadjective comparison (Suzuki, forthcoming), and a logic of questions and answers by means of measurement theory.

Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research (C) (23520015). Preliminary versions of this paper were presented at LOFT 2010 (Toulouse, France, (Suzuki, 2010)) and SOCREAL 2010 (Sapporo, Japan). The author would like to thank two anonymous reviewers of SOCREAL 2010 for their helpful comments.

References

Ackerman, F. (1994), 'Roots and Consequences of Vagueness', *Philosophical Perspectives*, 8: 129–136.

Armstrong, E. W. (1939), 'The Determinateness of the Utility Function', *Economic Journal*, 49: 453–467.

Boutilier, C. (1994), 'Toward a Logic for Qualitative Decision Theory', in *Proceedings of the 4th International Conference on Principles of Knowledge Representation and Reasoning (KR-94)*, Bonn, 75–86.

Cantor, G. (1895), 'Beiträge zur Begründung der Transfiniten Mengenlehre I', *Mathematische Annalen*, 46: 481–512.

Chisholm, R. M. and Sosa, E. (1966), 'On the Logic of Intrinsically Better', *American Philosophical Quarterly*, 3: 244–249.

Davidson, D. et al. (1955), 'Outline of a Formal Theory of Value, I', *Philosophy of Science*, 22: 140–160.

Fechner, G. T. (1860), *Elemente der Psychophysik*, Leipzig: Breitkopf und Hartel.

Fishburn, P. C. (1970), 'Intransitive Indifference in Preference Theory: A Survey', *Operations Research*, 18: 207–228.

Halldén, S. (1957), *On the Logic of 'Better'*, Lund: CWK Gleerup.

Hansson, B. (1968), 'Fundamental Axioms for Preference Relations'. *Synthese*, 18: 423–442.

Hansson, S. O. (2001), 'Preference Logic', in Gabbay, D. M. and Guenthner, F. (eds.), *Handbook of Philosophical Logic, 2nd Edition*, Vol. 4, Dordrecht: Kluwer, 319–393.

Hansson, S. O. and Grüne-Yanoff, T. (2006), 'Preferences', in *Stanford Encyclopedia of Philosophy*.

Huber, O. (1974), 'An Axiomatic System for Multidimensional Preferences', *Theory and Decision*, 5: 161–184.

Huber, O. (1979), 'Nontransitive Multidimensional Preferences: Theoretical Analysis of a Model', *Theory and Decision*, 10: 147–165.

Keefe, R. (2000), *Theories of Vagueness*, Cambridge: Cambridge University Press.

Lehrer, K. and Wagner, C. (1985), 'Intransitive Indifference: The Semi-order Problem', *Synthese*, 65: 249–256.

Luce, D. (1956), 'Semiorders and a Theory of Utility Discrimination', *Econometrica*, 24: 178–191.

Martin, R. M. (1963), *Intension and Decision*, Englewood Cliffs: Prentice-Hall.

Roberts, F. S. (1979), *Measurement Theory*, Reading: Addison-Wesley.

Schick, F. (1986), 'Dutch Bookies and Money Pump', *The Journal of Philosophy*, 83: 112–119.

Scott, D. (1964), 'Measurement Structures and Linear Inequalities', *Journal of Mathematical Psychology*, 1: 233–247.

Scott, D. and Suppes, P. (1958), 'Foundational Aspects of Theories of Measurement', *Journal of Symbolic Logic*, 3: 113–128.

Segerberg, K. (1968), 'Decidability of S4.1', *Theoria*, 34: 7–20.

Segerberg, K. (1971), 'Qualitative Probability in a Modal Setting', in Fenstad, J. E. (ed.), *Proceedings of the Second Scandinavian Logic Symposium*, Amsterdam: North-Holland, 341–352.

Simon, H. A. (1982), *Models of Bounded Rationality*, Cambridge, Mass.: The MIT Press.

Suzuki, S. (2009a), 'Prolegomena to Dynamic Epistemic Preference Logic', in Hattori, H. et al. (eds.), *New Frontiers in Artificial Intelligence*, LNAI 5447, Heidelberg: Springer-Verlag, 177–192.

Suzuki, S. (2009b), 'Measurement-Theoretic Foundation of Preference-Based Dyadic Deontic Logic', in He, X. et al. (eds.), *Proceedings of the Second International Workshop on Logic, Rationality, and Interaction (LORI-II)*, LNAI 5834, Heidelberg: Springer-Verlag, 278–291.

Suzuki, S. (2009c), 'Measurement-Theoretic Foundation of Logic for Goodness and Badness', in Bekki, D. (ed.), *Proceedings of the Sixth Workshop on Logic and Engineering of Natural Language Semantics (LENLS 2009)*, JSAI, 1–14.

Suzuki, S. (2010), 'Prolegomena to Threshold Utility Maximiser's Preference Logic', in *Electronic Proceedings of the 9th Conference on Logic and the Foundations of Game and Decision Theory (LOFT 2010)*, Paper No. 44.

Suzuki, S. (2011a), 'Prolegomena to Salient-Similarity-Based Vague Predicate Logic', in Onoda, T. et al. (eds.), *New Frontiers in Artificial Intelligence*, LNAI 6797, Heidelberg: Springer-Verlag, 75–89.

Suzuki, S. (2011b), 'Measurement-Theoretic Foundations of Probabilistic Model of JND-Based Vague Predicate Logic', in van Ditmarsch, H. et al. (eds.), *Proceedings of the Third International Workshop on Logic, Rationality, and Interaction (LORI-III)*, LNAI 6953, Heidelberg: Springer-Verlag, 272–285.

Suzuki, S. (forthcoming), 'Measurement-Theoretic Foundations of Interadjective-Comparison Logic', forthcoming in *Proceedings of Sinn und Bedeutung 16*.

Tversky, A. (1969), 'Intransitivity of Preferences', *Psychological Review*, 76: 31–48, rpt. in Shafir, E. (ed.) (2004), *Preference, Belief and Similarity: Selected Writings / by Amos Tversky*, Cambridge, Mass.: The MIT Press, 433–461.

Van Benthem, J. and Liu, F. (2007), 'Dynamic Logic of Preference Upgrade', *Journal of Applied Non-Classical Logics*, 17: 157–182.

Van Benthem, J. et al. (2005), 'Preference Logic, Conditionals and Solution Concepts in Games', ILLC Prepublication Series PP-2005-28.

Van Benthem, J. et al. (2009), 'Everything Else Being Equal: A Modal Logic for *Ceteris Paribus* Preferences', *Journal of Philosophical Logic*, 38: 83–125.

Van Rooij, R. (2011), 'Revealed Preference and Satisficing Behavior', *Synthese*, 179: 1–12.

Von Wright, G. H. (1963), *The Logic of Preference*, Edinburgh: Edinburgh University Press.

Notes to Contributors

1. All submitted papers are subject to anonymous peer-review, and will be evaluated on the basis of their originality, quality of scholarship and contribution to advancing the understanding of applied ethics.
2. Papers should not exceed 8,000 words including references.
3. Papers must be accompanied by an abstract of 150-300 words.
4. Submission should be made through e-mail to caep@let.hokudai.ac.jp
5. In-text references should be cited in standard author-date form: (Walzer 1977; Kutz 2004), including specific page numbers after a direct quotation, (Walzer 1977, 23-6).
6. A complete alphabetical list of references cited should be included at the end of the article in the following style:

   Walzer, M. (1977), *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, New York: Basic Book.

   Kutz, C. (2004), 'Chapter 14: Responsibility', in J. Coleman and S. Shapiro (eds.), *Jurisprudence and Philosophy of Law*, Oxford, UK: Oxford University Press, 548-87.

   Cohen, G.A. (1989), 'On the Currency of Egalitarian Justice', *Ethics*, 99 (4): 906-44.
7. Accepted papers will appear in both web-based electronic and printed formats.
8. The editorial board reserves the right to make a final decision for publication.