
PUBLIC POLICY MASTER THESIS

May 2021

How to favour argument diversity on online consultative platforms?

An experiment on the effect of exposure to other participants' arguments on the diversity of aspects tackled

Sophie de Rouilhan

Master's Thesis supervised by Lou Safra

Second member of the Jury: Coralie Chevallier

Master in Public Policy

Policy Stream: Public Administration

Abstract

On online consultative platforms - a type of digital democracy tool where citizens are asked to put forth arguments relative to a public policy project - what matters is not only the quality of the posted arguments but also their diversity. Especially, the pool of arguments collected is expected to tackle all the aspects of the project under consideration. The number of aspects tackled by the pool of arguments can be influenced by the design of the platform, especially by whether or not the arguments of other participants are made visible. Thus, in this thesis, we try to answer the following question: how does the visibility of the arguments of previous participants on consultative platforms impact the number of aspects tackled by the collected pool of arguments? Existing literature in psychology suggests that it should lead to a decrease in the level of aspect diversity achieved by groups, because of the will to respond to others' arguments, informational influence, normative social influence, or possibly a downward social comparison effect. To test this hypothesis, we designed an online experiment, whereby we asked participants to produce arguments as if they were on a consultative platform. Depending on the condition they were put in, they could either see four arguments or none. We compared, using a resampling technique, the probability that same-sized groups from the two conditions would tackle at least a certain number of aspects. The results show that exposure to arguments does tend to decrease the probability that groups achieve a high level of aspect diversity. Finally, we discuss directions for future research and possible implications of those results for platform designers.

Key words

digital democracy, e-participation, consultative platforms, argumentation, argument diversity, semantic categories

Table of content

WHY SHOULD I READ THIS RESEARCH?	4
INTRODUCTION	5
<u>1. WHY SHOULD WE CARE ABOUT ARGUMENT DIVERSITY ON ONLINE CONSULTATIVE PLATFORMS? A STATE OF KNOWLEDGE.....</u>	<u>9</u>
1.1. DEFINING ONLINE CONSULTATIVE PLATFORMS.....	9
1.2. DIVERSITY OF CONTRIBUTIONS AS AN ASSET FOR PARTICIPATORY AND CROWDSOURCING PLATFORMS.....	11
1.2.1. PARTICIPATORY AND CROWDSOURCING PLATFORMS HELP MAKING BETTER DECISIONS... ..	11
1.2.2. ... THROUGH THE DIVERSITY OF THE CONTRIBUTIONS THEY ALLOW TO COLLECT.	11
1.3. ARGUMENT DIVERSITY AS AN ASSET AND AS A GOAL FOR CONSULTATIVE PLATFORMS.....	13
1.3.1. SEVERAL CROWDSOURCING PLATFORMS DO COLLECT ARGUMENTS	13
1.3.2. ARGUMENT DIVERSITY IS USEFUL TO DECISION-MAKERS.....	14
<u>2. RESEARCH QUESTION AND HYPOTHESES: THE EFFECT OF THE EXPOSURE TO OTHER PARTICIPANTS' ARGUMENTS ON ASPECT DIVERSITY</u>	<u>21</u>
2.1. RESEARCH QUESTION.....	21
2.2. HYPOTHESES.....	22
2.2.1. MAIN HYPOTHESIS: WHEN OTHER PARTICIPANTS' ARGUMENTS ARE VISIBLE, ASPECT DIVERSITY DECREASES AT THE GROUP LEVEL	22
2.2.2. ALTERNATIVE HYPOTHESIS: WHEN OTHER PARTICIPANTS' ARGUMENTS ARE VISIBLE, ASPECT DIVERSITY INCREASES AT THE GROUP LEVEL	23
<u>3. MATERIALS AND METHODS.....</u>	<u>24</u>
3.1. PARTICIPANTS	24
3.1.1. RECRUITMENT	24
3.1.2. SAMPLE SIZE RATIONALE	25
3.1.3. INCLUSION CRITERIA CONCERNING PARTICIPANTS' PROFILE	25
3.2. PROCEDURE.....	25
3.2.1. THE TWO CONDITIONS	26
3.2.2. THE ARGUMENTATIVE TASK.....	26
3.2.3. ATTENTION CHECK QUESTIONS.....	28
3.2.4. QUESTIONS ABOUT INDIVIDUAL CHARACTERISTICS	29
3.3. DATA ANALYSIS	29
3.3.1. DATA EXCLUSION.....	30
3.3.2. DEPENDENT VARIABLES	30
3.3.3. CODING METHODOLOGY.....	32
<u>4. RESULTS</u>	<u>35</u>
4.1. PILOT STUDY.....	35
4.1.1. EXCLUSION CRITERIA	35
4.1.2. CODING RELIABILITY.....	36

4.2. MAIN STUDY	39
4.2.1. COMPARISON BETWEEN THE TWO PARTICIPANT SAMPLES (A AND B)	39
4.2.2. CODING RELIABILITY	41
4.2.3. STATISTICS ABOUT THE ARGUMENTS COLLECTED	45
4.2.4. DEPENDENT VARIABLES	46
5. DISCUSSION AND RECOMMENDATIONS.....	55
5.1. DISCUSSION REGARDING THE ECOLOGICAL VALIDITY OF THE STUDY: WHAT CAN OUR RESULTS TELL US ABOUT REAL CONSULTATIVE PLATFORMS?	55
5.2. RECOMMENDATIONS FOR DESIGNING CONSULTATIVE PLATFORMS	56
5.3. DIRECTIONS FOR FUTURE RESEARCH: WHAT DO PLATFORM DESIGNERS STILL NEED TO KNOW?	58
CONCLUSION.....	61
BIBLIOGRAPHY	62
APPENDIX.....	67
APPENDIX I. INSTRUCTIONS	67
APPENDIX II. ATTENTION CHECK QUESTIONS ABOUT URBAN TOLLS.....	70
APPENDIX III. ARGUMENTS FAILING THE RELEVANCE CHECK.....	71
APPENDIX IV. LIST OF ASPECTS	72
APPENDIX V. PERCENTAGE OF ARGUMENTS TACKLING EACH ASPECT	73
APPENDIX VI. CODING RELIABILITY FOR THE MAIN STUDY BEFORE CORRECTION OF THE CARELESS MISTAKES.....	74
APPENDIX VII. COMPARISON BETWEEN $P_{N.A}(Y_i \geq X)$ AND $P_{N.B}(Y_i \geq X)$ WITHOUT OVERLAPPING A-GROUPS	75
APPENDIX VIII. REGRESSION RESULTS FOR INDIVIDUAL-LEVEL VARIABLES	76

Why should I read this research?

For today's policymakers, it is often essential to involve citizens in their decision process. Indeed, involving citizens has numerous advantages that can be decisive in making our democracies resistant and efficient: it can improve trust in public officials and in democratic institutions, it can increase the legitimacy of decisions, and, last but not least, it can help making better and more efficient decisions. Because digital tools, such as online platforms, allow to reach more citizens than ever before, while being comparatively inexpensive and easy to set up, they appear particularly well-suited to develop a broad and efficient citizen participation.

However, it is not enough to make it possible for citizens to be involved in the decision process to harvest all the benefits of citizen participation. Indeed, the participation space must be carefully designed so as to induce the desired participatory behaviour, otherwise the policymakers might end up collecting inputs that are not at all those they wanted to obtain. This is *a fortiori* the case for online participatory platforms. If platforms are poorly designed, citizens may either be discouraged from participating or may fail to produce the type of contribution that is wanted. There is no need to say that the stakes are high: a poorly designed platform will not only fail to generate all the positive effects of efficient participation, it will have wasted resources, and it may even lead to a higher distrust or disinterest from citizens, who will feel like their efforts to participate resulted in nothing. This is why knowing precisely how design features of online platforms influence participants' behaviour is of crucial importance.

Numerous studies have already been made regarding how design could help obtain diverse desirable results on online participatory platforms. However, a certain type of participatory platform, which we call here *consultative platforms*, has been overlooked in existing literature. Especially, one of its major goals, which is to collect a pool of arguments that exhibits a high level of diversity, has been totally ignored. As a consequence, to our knowledge, no study has been made regarding how the design features of such platforms can influence the level of diversity of the argument pool they collect.

The goal of this thesis is to start filling this gap. Drawing mostly on literature from the cognitive sciences (especially on brainstorming, argumentation and group decision-making) we identified a design feature that was particularly likely to have an impact on argument diversity. This design feature is the visibility of other participants' arguments on platforms. In this thesis, we studied the impact of this design feature on one specific dimension of argument diversity, namely the number of aspects of the issue that are tackled by the argument pool. We tested this impact through an online experiment, and found that exposure to others' arguments tended to decrease the level of aspect diversity of the argument pools produced by groups. Of course, further research will be needed to replicate those results before they can be considered robust. However, if those results were replicated, it would mean that several designers of consultative platforms may have to adapt their design strategies if they want their platforms to collect arguments that cover all the aspects of the problem at hand.

Introduction

In recent years, digital tools have been increasingly used for democratic purposes. At a time when distrust in governments and general loss of confidence in democratic institutions increase, at a time also when the value of “citizen expertise” and citizen input in policymaking is more and more recognized, digital tools appear as a particularly fruitful and innovative way to both fight distrust and tap into citizen expertise, by involving citizens in policymaking processes (Simon *et al.*, 2017). There are numerous ways to involve citizens, this is why numerous types of digital tool have been invented: for instance participatory budget platforms, debating platforms, collaborative documents (Simon *et al.*, 2017). In this thesis, we focus on a particular type of digital tool, which we call *consultative platforms*. Consultative platforms are online platforms on which citizens are asked by public officials to spell out their reasons to support or oppose a public policy project, but are not given any decisional power. Concretely, it means that consultative platforms collect *arguments* about projects. The term “argument” is indeed defined by online Cambridge dictionary as a “reason” given to “support or oppose an idea or suggestion”. We only make one addition to this definition: on consultative platforms, the reasons for supporting an idea must be based on characteristics of this idea, and not on the mere idiosyncrasy of the citizen. In other words, “I am against this idea because I am in a bad mood today” would not qualify as a (relevant) argument in a consultative platform. Thus, when we talk about arguments later on, we do not include such irrelevant arguments.

In France, consultative platforms have been launched for instance regarding a project of universal income (*Revenu Universel d’Activité*, in French) in 2019 (<https://www.consultation-rua.gouv.fr/>), or regarding a possible reform of the civic service (*service civique*) in 2020 (<https://consultation.service-civique.gouv.fr/>). On online platforms, design possibilities are virtually limitless, and each design choice can have a significant impact on the behaviour of participants. Thus, it is of crucial importance to know how specific design choices impact behaviour, so as to make sure that the platform will best fulfil its goal(s). Several studies (e.g. Towne and Herbsleb, 2012; Aitamurto and Landemore, 2015) already exist on how design choices can help achieve certain results on online platforms (such as increasing the usability of the platform, attracting more participants, or increasing the quality of contributions). However, one of the goals of consultative platforms has been largely ignored by the literature so far: namely, that of collecting arguments that are sufficiently *diverse*. As a consequence, to our knowledge, no study has yet been made about how design choices could help achieve this particular goal.

There is actually a large awareness in the literature on online platforms used for policy making (e.g. Towne and Herbsleb, 2012; Aitamurto and Chen, 2017; Taeihagh, 2017) that the diversity of inputs collected on platforms is beneficial for policymakers, in that it helps them come to a better decision (i.e. a decision that has higher positive effects and/or a lesser cost and/or smaller negative side-effects than the other possible options) by furnishing them with supplementary data. Importantly, diversity can mean two things in this context. Firstly, it can refer to the mere fact that the contributions are different, i.e. non-redundant. Secondly, it can refer to the fact that the contributions tackle different *aspects* of the public policy under

consideration (for instance, its cost, its targeted public, its environmental consequences, etc.). These two meanings can be seen as two dimensions along which diversity can be evaluated. When we talk generically of “contribution diversity” on a platform, whether it is *argument diversity*, *idea diversity* or *information diversity*, we refer to both those dimensions conjointly, i.e. to the total number of non-redundant contributions collected (whether they are arguments, ideas, or pieces of information) *and* to the number of aspects tackled by these contributions altogether.

Up to now, however, the literature on platforms has focused on the usefulness of increasing idea diversity (e.g. Simon *et al.*, 2017) and information diversity (e.g. Aitamurto and Chen, 2017), but has ignored *argument* diversity. We argue that there is no good reason for this omission, because what is true for ideas and information is also true for arguments: the higher the number of non-redundant arguments collected and the more the arguments explore the different aspects of the issue at hand, the higher the probability that the pool of arguments will help policymakers make a good decision. Actually, existing literature in cognitive and behavioural sciences (especially on cognitive diversity, reasoning and group decision-making) provides strong evidence that collecting a very diverse pool of arguments can indeed be useful to policymakers. Indeed, groups of policymakers are likely to lack cognitive diversity (Landemore, 2013), which may prevent them from finding easily very diverse arguments. They are also likely to come to an early consensus due to groupthink effects (Sunstein and Hastie, 2015), which may lead them to prematurely stop looking for arguments (this is especially suggested by the argumentative theory of reasoning: Mercier, 2016) and thus would prevent them from considering a large pool of arguments. Thus, policymakers may fail to produce by themselves a pool of arguments that achieves a high level of diversity, which is why they can benefit from collecting such a pool through consultative platforms.

Once we have established that increasing argument diversity (in both its dimensions) is indeed a desirable outcome for consultative platforms, we are faced with the following question: how should platforms be designed so as to maximize argument diversity? On this point, as noted above, information is lacking. As research has focused until now on the importance of idea diversity and information diversity, there has not yet been, to our knowledge, any study on how to increase argument diversity. This is why, in this thesis, our objective is to start exploring the question of how consultative platforms can, through the design they use, enhance arguments diversity.

In order to determine what design choices can have an impact on argument diversity, we draw on existing literature in cognitive sciences, especially concerning brainstorming, group decision-making, and argument production. This literature suggests that the exposure to other participants’ arguments is very likely to have a significant impact on argument diversity, and especially on the dimension of diversity that is relative to the aspects tackled by the arguments collected on the platform (i.e. the second dimension of argument diversity). Moreover, it is interesting to note that platforms largely differ on whether or not others’ arguments are visible to participants. On some platforms, existing contributions are immediately visible to new participants (ex: several platforms designed by Cap Collectif, such as this one :

<https://www.consultation-rua.gouv.fr/>). On other platforms, people write their contributions without being able to access what others have written (ex: several platforms designed by Delib, such as this one: <https://consult.gov.scot/>). This seems to confirm that the visibility of others' arguments is a controversial design choice among platform designers. Thus, investigating the impact of this design choice can help platform designers make more informed decisions. Hence, the objective of this thesis is to answer the following question: **what is the impact of seeing other participants' arguments on aspect diversity in consultative platforms?**

The reviewed literature in cognitive sciences suggests that exposure to others' arguments should lead to a decrease in aspect diversity. Indeed, studies on idea diversity in brainstorming tasks have found that exposure to others' *ideas* can lead to a decrease in the number of aspects tackled by the pool of ideas produced by a group (Nijstad, Bechtoldt and Choi, 2019). Besides, several social and cognitive processes which are likely to underlie argument production, could increase such an effect. Indeed, people could want to respond to specific arguments, which would make them more likely to tackle the same aspects as those arguments (this is especially suggested by the argumentative theory of reasoning: Mercier, 2016). Besides, participants could be subject to a normative social influence, whereby seeing others' arguments would incite them to tackle the same aspects as others for reputational purposes (such a phenomenon has been observed for information sharing: Wittenbaum, Hubbell and Zuckerman, 1999; Wittenbaum and Park, 2001). Participants could also be subject to an informational influence, in which case seeing others' arguments could change their opinion on the importance and relevance of certain aspects, which would induce them to tackle those aspects (informational influence has been observed in decision-making groups : Sunstein and Hastie, 2015). An alternative hypothesis, however, would be that exposure to others' arguments would lead to an increase in aspect diversity, due to cognitive stimulation and/or social stimulation. Those phenomena have been shown to increase the number of non-redundant ideas produced in brainstorming tasks (e.g. Yagolkovskiy, 2016; Fink *et al.*, 2012; Leggett Dugosh and Paulus, 2005), but they have not been shown to increase the number of *aspects* tackled. Thus, this hypothesis is less likely.

In order to test these hypotheses, we designed an online experiment, whereby we asked participants to produce arguments on a public policy project, namely the introduction of an urban toll in their city, as if they were on a real consultative platform. In this experiment, two conditions were distinguished. In *condition A*, participants didn't have access to any argument, as if they were on a consultative platform where previous contributions are not visible. In *condition B*, on the other hand, participants could read four arguments of previous participants. This experiment is meant to evaluate first and foremost a collective effect: we want to know how seeing others' arguments impacts the number of aspects tackled by the pool of arguments that a *group* of participants produces. Indeed, what matters in a consultative platform is not that each participant tackles *individually* a high number of aspects but that the group of participants tackles *collectively* as many aspects as possible. Thus, in our experiment, we compare the probability that same-sized groups of both conditions would tackle a certain number of aspects. We also measure the effects of seeing others' arguments at an individual level, in order to collect some complementary data and to better understand what individual-level effects might

explain the collective effect demonstrated. In order to measure those variables, we developed an original coding scheme to analyse the arguments collected. The process of elaborating this coding scheme included testing it and improving it through a pilot experiment.

Our results validate our main hypothesis: exposure to others' arguments decreases the level of aspect diversity achieved by groups. However, they do not provide definitive evidence as to what individual behaviours underlie this collective effect. Of course, our experiment alone is not sufficient to prove that exposure to others' arguments systematically leads to a decrease in aspect diversity at the group-level. Future research is needed to see whether those results can be replicated in different contexts. However, if our results were to be confirmed, then it would have important consequences regarding how to best design consultative platforms.

In the first section of this thesis, mainly through a review of existing literature, we specify what consultative platforms are and we explain why we should care about argument diversity on such platforms. In the second section, we expose our research question and our hypotheses. In the third section, we detail the experimental design and the methodology used for data analysis. In the fourth section, we present the results of our pilot study and of our main study. Finally, in the fifth and last section, we discuss the ecological validity of our study and what recommendations can be made regarding platform design and directions for future research.

1. Why should we care about argument diversity on online consultative platforms? A state of knowledge.

1.1. Defining online consultative platforms

Digital tools and technologies have been increasingly used for democratic purposes in recent years. This practice is called “digital democracy” (Simon *et al.*, 2017). A very large literature exists on digital democracy tools, and numerous concepts are used to define them, which more or less overlap with one another: “digital participatory platforms” (Falco and Kleinhans, 2019), “online civic engagement platforms” (Nelmarkka *et al.*, 2014), “open government platforms” (Koch, Füller and Brunswicker, 2011), etc. Each of these large categories is then divided in sub-categories to produce a finer-grained typology (see for instance Simon *et al.*, 2017; Nelmarkka *et al.*, 2014; Falco and Kleinhans, 2019).

Our focus is on a certain subcategory of what has been called “*crowdsourcing platforms*”. Importantly, crowdsourcing can be used in diverse realms (Aitamurto and Chen, 2017) but only crowdsourcing used for policymaking is of interest here. Thus, in this thesis, the phrase “crowdsourcing platforms” only refers to platforms used for crowdsourced policymaking. In this sense, crowdsourcing platforms are a sub-category of “digital participatory platforms” (as defined by Falco and Kleinhans, 2019), that is to say platforms which are explicitly built to create a link between citizens and governments, as opposed to other digital tools which can be used to establish this link but have not been designed explicitly for this purpose (such as Twitter). More precisely, crowdsourcing consists in “an open call for anybody to participate in an online task” (Aitamurto and Chen, 2017), where the task generally consists in a “one-time contribution that does not involve working with other contributors” (Aitamurto and Landemore, 2015). Crowdsourcing platforms aim thus at collecting citizens’ contributions concerning a policy project, but not at involving citizens in a long-term teamwork or process. Another important feature of crowdsourced policymaking is that it does not imply conferring any decisional power to citizens (Aitamurto and Chen, 2017).

On crowdsourcing platforms, citizens’ contributions can take many different forms (votes, ideas, information, etc.). Our focus is on crowdsourcing platforms which collect contributions under the form of *arguments about public policy proposals*. Importantly, it does not matter whether the platform’s purpose is to collect argument or if the platform has a different purpose. It does not matter either whether the platform allows for other types of contributions besides arguments (most platforms contain several functionalities and allow for different types of contributions). The platforms we study in the following sections are defined only by two characteristics: they are crowdsourcing platforms, and they allow to produce arguments concerning public policy proposals. As existing literature has not proposed a specific concept to designate those platforms, for simplicity’s sake, we will call them from now on

consultative platforms. This terminological choice is justified by the fact that the term of consultation is the one usually used when people are asked for their opinions and arguments about something, without being asked to produce long-term work or to make decisions. Actually, the notion of consultation is the one used by several firms which design those platforms: the firm Delib refers to its consultative platforms as “consultation hubs” while Cap Collectif talks of “consultation platforms” (in French: “*plateformes de consultation*”). Examples of platforms designed by those two firms are given later on.

To sum up, consultative platforms:

1. are built with the explicit purpose of being a digital democracy tool, that is to say of creating a link between public decision-makers and citizens.
2. are open to any citizen willing to participate. This excludes for instance representative polls happening online, where only a predetermined representative sample of citizens would be allowed to participate.
3. are meant to offer the possibility of a one-time contribution, not to involve citizens in a long-term teamwork or process.
4. give citizens no decisional power.
5. allow citizens to produce *arguments* concerning a public policy proposal (possibly among other types of contributions, and possibly for diverse purposes).

In spite of the large literature on digital democracy tools, there has been no specific focus on consultative platforms, and there is no concept that has been devised to capture them specifically as a significant sub-category. Most of the time, the functionality of allowing argument production is seen as one way among others to obtain something else such as information, ideas or preferences. The fact that the preferences, information, or whatever else is wanted, take the form of arguments in some of these crowdsourcing platforms is ignored (see for instance the “*typology of digital democracy*” developed by Simon *et al.*, 2017). In the cases where argument production is specifically targeted, it is exclusively associated with deliberative ideals and purposes, that is to say, with the ideal of an exchange of arguments among free and equal individuals (see for instance Aitamurto and Landemore, 2016, and Aitamurto, 2016, for a review of the goals and ideals associated with deliberation and a reflection on the type of deliberation created by argument production on crowdsourcing platforms). As a consequence, argument production is only studied on crowdsourcing platforms allowing for some kind of deliberation between participants, leaving aside platforms allowing for argument production but not for deliberation (the typical example being the Consultation hubs designed by the firm *Delib* on which citizens cannot see other participants’ arguments, which renders any kind of deliberation between citizens impossible). As we shall argue hereafter, the fact that there has been no specific focus on consultative platforms has resulted in the literature on crowdsourcing platforms omitting one important goal of many of those platforms: collecting diverse arguments.

1.2. Diversity of contributions as an asset for participatory and crowdsourcing platforms

Let us, first of all, deal with a terminological issue. In the following section, we use some articles which mostly focus on crowdsourcing platforms but consider them only under some broader concept, such as “digital participatory platform”, without explicitly distinguishing them. That is the case of Brabham and Guth (2017); R. Farina *et al.* (2013) ; Simon *et al.* (2017); Towne and Herbsleb (2012). As a consequence, the conclusions of those articles do apply to crowdsourcing platforms even if they are not explicitly identified as such. Thus, to avoid confusion, we use the phrase “participatory and crowdsourcing platforms” when the articles we refer to do not all use the concept of crowdsourcing platform.

1.2.1. Participatory and crowdsourcing platforms help making better decisions...

Participatory platforms can have diverse purposes: increasing the legitimacy of the decision, kindling citizens’ interest in political matters, etc. Existing literature shows that one of these purposes is helping the decision-makers to make better decisions. Brabham and Guth (2017) interviewed founders and executives of firms designing participatory platforms, and found that one of the “ideals” of founders and executives was that their platform would lead to “improved problem-solving and decision-making in government”. Not only are many platforms designed with the explicit purpose of improving decision-making, but existing literature on participatory and crowdsourcing platforms argues that such improvement does indeed happen (see for instance Towne and Herbsleb, 2012; R. Farina *et al.*, 2013; Simon *et al.*, 2017; Aitamurto and Chen, 2017; Taeihagh, 2017).

1.2.2. ... through the diversity of the contributions they allow to collect.

The improvement in decision-making produced by participatory and crowdsourcing platforms is attributed to the fact that they allow to increase the pool of ideas, information, experiences and perspectives taken into account in the decision-making process (Towne and Herbsleb, 2012; R. Farina *et al.*, 2013; Simon *et al.*, 2017; Aitamurto and Chen, 2017; Taeihagh, 2017). As a consequence, collecting *diverse* contributions appears clearly as an asset for these platforms. The more diverse the contributions collected by a platform, the larger the pool of ideas, information, or else made accessible to the decision-makers, and thus the higher the chance that this platform will lead to a better decision.

Importantly, diversity has two different dimensions. Firstly, it can refer to the mere fact that contributions are different, i.e. non-redundant. Secondly, it can refer to the fact that contributions tackle different *aspects* of the public policy project under consideration. The term “aspect” here refers to the general features of the public policy, such as its attributes, its consequences, or other features it may depend on. For instance, regarding a project of urban toll, a feature of the policy could be the rate of the toll, a consequence could be the impact of the toll on pollution, and a feature it depends on could be the decision process which will lead to its possible introduction. To each of the two dimensions of diversity can correspond a measure: to the first dimension corresponds the measure of the number of non-redundant contributions in a pool of contributions; to the second dimension corresponds the measure of the number of aspects tackled in a pool of contributions, which corresponds to measuring the extent to which the pool of contributions has explored the totality of the “problem-space”. Although in literature about crowdsourcing platforms there is no distinction between those two dimensions of diversity, the usefulness of both has been underlined by the literature on idea brainstorming (Paulus and Kenworthy, 2019), albeit not with the same terminology. In literature about idea brainstorming, aspects are referred to as “semantic categories”. Moreover, various terms are used to refer to the two dimensions of diversity. The number of non-redundant ideas produced can be referred to for instance as “idea fluency”, “fluidity” or “productivity” (Nijstad, Stroebe and Lodewijkx, 2002; Dennis, Minas and Williams, 2019), while the number of semantic categories tackled can be referred to for instance as “diversity” or “flexibility” (Ziegler, Diehl and Zijlstra, 2000; Nijstad, Stroebe and Lodewijkx, 2002). Expressions such as “idea fluency”, “fluidity”, “productivity” characterize the *process* of producing ideas. Because our focus is not on the process but on the pool of ideas (or other contributions) produced, we choose to use the term “diversity”.

On participatory and crowdsourcing platforms, both dimensions of diversity are useful for decision-makers. Indeed, if decision-makers need to collect contributions on a particular problem (because they don’t have enough ideas or information, or else), it should be useful for them to obtain not only a high number of non-redundant contributions in total, but also to collect a pool of contributions that explores all the problem-space, i.e. that tackles all the relevant aspects of the issue at hand. Thus, *a priori*, the most reasonable assumption is that both types of diversity on platforms are equally useful, and that the best option is, if possible, to maximize both. Thus, as both dimensions of diversity are *a priori* equally useful (and since they are not mutually exclusive), when we use the notion of “diversity” without any precision in the following, we refer to both dimensions conjointly.

The literature on participatory and crowdsourcing platforms focuses mainly on the usefulness of collecting diverse ideas (e.g. Simon *et al.*, 2017) or diverse pieces of knowledge (e.g. Aitamurto and Chen, 2017). The question of the usefulness of collecting diverse arguments is ignored, even though many crowdsourcing platforms ask citizens not only for ideas or information but also for arguments. This is probably due to the fact that, as explained above, existing literature has either overlooked the specific functionality of allowing for argument production or has associated it exclusively with deliberative purposes. Actually, there are good

reasons to believe that argument diversity is at least as useful for crowdsourcing platforms as idea diversity or information diversity. In the following section, we detail those reasons.

1.3. Argument diversity as an asset and as a goal for consultative platforms

There are two main reasons why argument diversity should be considered as an asset for participatory platforms and should be studied accordingly. Before reviewing those reasons, we need to briefly explain why arguments are not equivalent to ideas or knowledge (if they were, our whole argumentation would of course be moot).

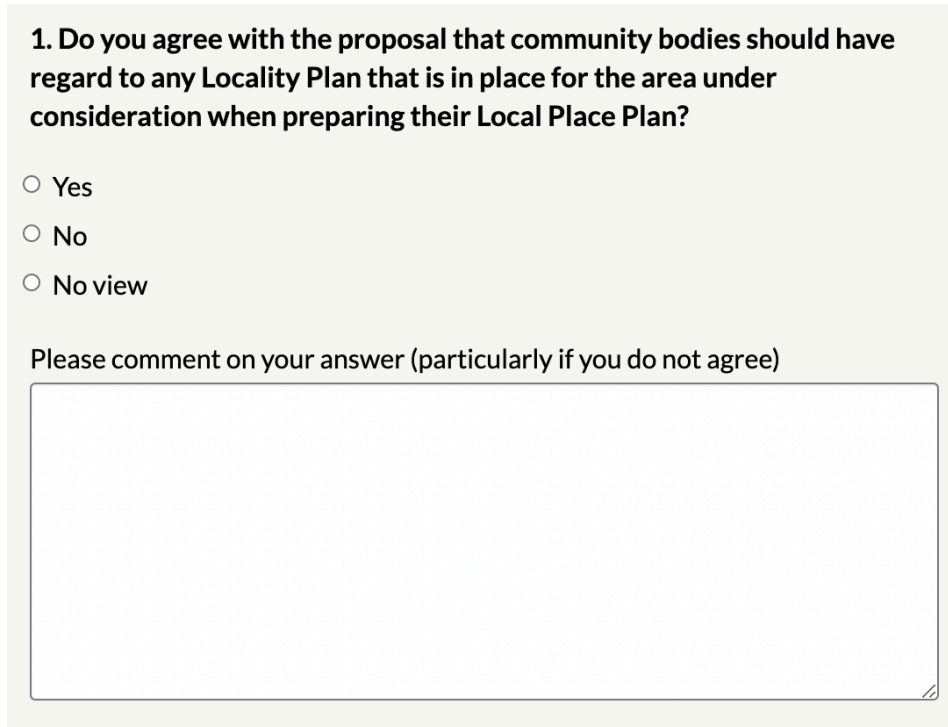
An idea corresponds to a possible option. For instance, in idea brainstorming tasks, participants are asked to find possible ways to improve tourism (Gallupe *et al.*, 1992), possible uses for a knife (*ibid*), or possible solutions to help preserve the environment (Nijstad, Stroebe and Lodewijkx, 2002). An argument, on the other hands, consists of a reason why a certain option is or is not interesting/feasible/etc. The difference between ideas and arguments is acknowledged in the literature about idea brainstorming: actually, one of the central questions in this field is whether isolating the task of producing ideas from the task of producing evaluations of those ideas (i.e. from producing *arguments* concerning those ideas) increases the efficiency of the brainstorming (Gallupe *et al.*, 1992). As for knowledge, whether the latter is expertise-based or experience-based, it consists basically of pure (factual) information. Of course, a piece of information can be used as a basis for an argument, but the piece of information in itself (i.e. the raw uninterpreted fact) is not an argument (the difference between knowledge and arguments is emphasized by Aitamurto, 2016). Thus, knowledge, ideas and arguments are three different types of input, and as a consequence a platform may wish to collect one of them without wishing to collect the others. Why platforms may wish to collect diverse pieces of knowledge or diverse ideas has been explained elsewhere (e.g. R. Farina *et al.*, 2013; Simon *et al.*, 2017; Aitamurto and Chen, 2017). In the following sections, we will try to show that platforms may also need and wish to collect diverse *arguments*.

1.3.1. Several crowdsourcing platforms do collect arguments

First of all, even if the one and only goal of platforms was to collect ideas or pieces of knowledge, platforms often collect them *under the form of arguments* (in other words, there is a large proportion of consultative platforms among crowdsourcing platforms). A good example of this is the consultation hubs designed by the firm *Delib*. They have been used by many governments and public officials around the world, for instance the Australian government (<https://consultations.health.gov.au/>) or the Scottish government (<https://consult.gov.scot/>; for more examples, see: https://www.delib.net/who_uses_delib). Consultations take the form of a series of questions asked to the participant, each followed by a blank square where people

elaborate on their answer. The questions often ask for citizens' opinions and arguments. An example is given below (figure 1).

Figure 1. A question asked as part of the “Local Place Plan Regulations” consultation launched by the Scottish Government in 2021 (on <https://consult.gov.scot/>)



1. Do you agree with the proposal that community bodies should have regard to any Locality Plan that is in place for the area under consideration when preparing their Local Place Plan?

☐ Yes

☐ No

☐ No view

Please comment on your answer (particularly if you do not agree)

Participants do not have the possibility to see others' answers and arguments, thus it appears clearly that asking participants to give reasons (i.e. arguments) for their opinions does not correspond to the will to create a deliberation of any kind. As a consequence, only two options remain: either the platform designers and public officials wanted to collect arguments, or they wanted to collect another type of input but decided to do so while asking for arguments. In either case, the cognitive processes underlying argument production (and not those underlying factual narration, idea brainstorming, or else) will be activated in the participants' mind, and whether those cognitive processes lead participants to propose very diverse arguments or not is of crucial importance, independently of the type of inputs the platform designers wanted to collect.

1.3.2. Argument diversity is useful to decision-makers

The second reason why argument diversity should be considered as an asset for crowdsourcing platforms is that, actually, it is useful in itself for decision-makers. Indeed, existing literature convincingly shows that collecting a diverse pool of arguments should in many cases help decision-makers improve their decision, as we shall now try to demonstrate.

As underlined by Manin (2004), “political and social decisions present in general multiple aspects and consequences. [...] As a consequence, during deliberation, each person can put forth, without repetition, different reasons for adopting the same policy”¹. In other words, political decisions can be supported (or undermined) by a very large number of different arguments. Indeed, in complex decisions, in which many parameters have to be taken into account, there is not the one and only “best argument” which will put an end to the debate and establish beyond doubt the superiority of one option. For each option, decision-makers will need to establish a long list of pros (arguments for the option) and cons (arguments against the option) concerning all the aspects of that option – i.e. all its attributes (its cost, its intended public, ...), its consequences (economic, social, ...), and other features it depends on –, to be able to properly evaluate its relative advantages. In other words, if an option is indeed superior to the others, many arguments (all the pros and cons on all the aspects of this options and the other possible ones) will be needed to prove it. Thus, there is indeed a large pool of arguments to be found, *and* there are indeed multiple aspects on which arguments need to be found. Having many arguments about one aspect gives a series of pros and cons relative to one parameter, while having arguments about each aspect gives pros and cons that cover the totality of the problem-space. As a consequence, in order to reach a good decision, decision-makers must include in their reflection arguments that tackle all the aspects of the decision, and the larger the number of non-redundant arguments relative of each aspect, the better. Obtaining a diverse pool of arguments on consultative platforms would not be such an asset for decision-makers, if they could, through intelligence, expertise, training and hard work, produce it themselves. But in many cases they will not be able to do so. Indeed, as we shall now try to show, several cognitive processes tend to limit the diversity of arguments on which decision-makers, whether in group or alone, base their decisions.

First of all, governmental decision-making groups (not to speak of single decision-makers) can lack cognitive diversity. Cognitive diversity refers to “the ability to see the world from different points of view” (Landemore, 2011). It refers more specifically to “a diversity of perspectives (the way of representing situations and problems), diversity of interpretations (the way of categorizing or partitioning perspectives), diversity of heuristics (the way of generating solutions to problems), and diversity of predictive models (the way of inferring cause and effect)” (Landemore, 2013). Landemore (2013), relying on results from Hong and Page (Hong and Page, 2001, 2004; Page, 2008), argues that cognitively diverse groups are better at problem-solving and deliberating. While her argumentation focuses mainly on the fact that cognitive diversity allows to pool more ideas and information, it seems highly probable that a cognitively diverse group is not only more likely to think of more ideas and information, but also more likely to think of more *arguments*, and probably also to consider a larger number of aspects. As a consequence, a single person or a group with a low level of cognitive diversity will probably conceive a less diverse pool of arguments than a group of people with a high level of cognitive diversity (at least, as long as the two groups are both reasonably competent about the issue and spend about the same amount of time and energy thinking of arguments). Governmental

¹ Translation from French: “*Les décisions politiques et sociales présentent en général des aspects et des conséquences multiples. [...] Dans la délibération chacun peut donc faire valoir, sans répétition, des raisons différentes d’adopter une même politique*”.

decision-making groups may lack cognitive diversity either because they are cognitively very similar (due to similar education, similar social background or else), but also simply because they involve a small number of people. Indeed, as argued by Landemore (2013), cognitive diversity should be highly correlated with the number of people in the group. In other words, including a much larger number of people in the decision process should automatically ensure a higher cognitive diversity, and, reversely, including only a small number of people will limit significantly cognitive diversity. Thus, small groups of decision-makers, especially if they are cognitively very similar, are likely to produce a not very diverse pool of arguments, which may be insufficient to find out the best possible decision.

One could still argue that this initial tendency to conceive a not very diverse pool of arguments could be counterbalanced by a long and thorough reflection aimed at finding new arguments. To some extent, it could indeed: it is not because people have a tendency to think of one kind of arguments that they cannot, through mental exertion, manage to find at least *some* new arguments tackling new aspects. However, it seems reasonable to think that one person or a few people are still unlikely to find, in a reasonable time, as much different arguments as a hundred people. Moreover, other mental processes limit the number of arguments decision-makers may take the time to conceive. Especially, the argumentative theory of reasoning, developed by H. Mercier and D. Sperber (Mercier and Sperber, 2011; Mercier, 2016), predicts that people are cognitively lazy and biased when producing arguments. This means two things. Firstly, that people don't look for more or better arguments than is necessary to convince others. The search for argument is thus "satisficing" (Mercier, 2011): people content themselves with finding a "good enough argument" (i.e. an argument that has a chance to convince others). They don't bother finding the best, most rigorous argument. Secondly, people don't think by themselves of the arguments that can be raised against their own position (Mercier, 2016). There is empirical evidence to support the predictions of the argumentative theory of reasoning: see for instance Toplak and Stanovich (2003) for people's tendency not to think of counterarguments, and see Trouche *et al.* (2016), for people's tendency to be satisfied with a "good enough" argument (and for a general review of evidence supporting the argumentative theory, see Mercier, 2016; Mercier *et al.*, 2017). As a consequence of such laziness and bias, in decision-making groups, as soon as a consensus is found, people are likely to stop looking for arguments (not to speak of single decision-makers, who, logically, are even less likely to bother finding multiple arguments). Thus, if a consensus is found very quickly, little or no counterargument will be taken into account, and the decision will only be based on the first arguments that came to mind (i.e. very few argument and not necessarily very good ones). This is likely not only to limit the total number of arguments but also possibly the number of aspects taken into consideration: indeed, the decision-makers may stop looking for arguments before having properly considered all the relevant aspects of the problem, especially aspects that would be tackled as part of counterarguments.

This is all the more problematic as groupthink effects can lead to a premature consensus among decision-making groups. Groupthink has been defined as "the psychological drive for consensus at any cost that suppresses disagreement and prevents the appraisal of alternatives in cohesive decision-making groups" (Ottaviani and Sørensen, 2001). Sunstein and Hastie (2015)

have linked groupthink especially to reputational and informational cascade. A cascade happens when “participants ignore their private knowledge and rely instead on the publicly stated judgement of others”(Sunstein and Hastie, 2015). A cascade is informational when the judgement of others is considered as information (“if those who spoke before me said that option A is the best option, they must have very good reason to think that. I deduce from that that option A is likely to be the best option, even though I had doubts about it”). A cascade is reputational when the judgement of others is considered as a social pressure (“if those who spoke before me said that option A is the best, I will suffer reputational damage if I disagree, so I will publicly agree, even though I may think privately otherwise”). Cascades tend to prevent debate and opposition, and to create a premature consensus. As a premature consensus would lead decision-makers to stop looking for arguments (following the argumentative theory of reasoning), this means that decision-makers may end up basing their decision on very few arguments, which do not necessarily tackle all the relevant aspects of the project.

Thus, if we sum up, decision-makers, due to lack of cognitive diversity, are unlikely to conceive initially a diverse pool of arguments. Due to cognitive laziness and myside bias (Mercier, 2016), if they come to an early consensus, they are unlikely to bother looking for more (or better) arguments, and especially they are unlikely to conceive any counterargument. And due to cascade effects, some groups are likely to come to an early consensus. All those cognitive processes converge to limit the number of arguments on which decision-makers base their decision, which is likely to lead to sub-optimal decisions. Those mental processes are also likely to lead decision-makers to focus on certain aspects of the problem while omitting of other important ones. Even if decision-makers do manage to tackle all the aspects of the problem in their reflection, they might not produce enough argument per aspect to really consider properly all aspects. Thus, decision-makers are likely to benefit not only from collecting more arguments, but from collecting more arguments on all the aspects of the problem. As a consequence, collecting diverse arguments (i.e. both a large number of non-redundant arguments and arguments that tackle all or most of the aspects of the problem) through consultative platforms appears to be a valuable purpose. Moreover, collecting diverse arguments should also help produce a more *balanced* pool of arguments, i.e. arguments for and against most points of views existing on most aspects of the subject (as long as the platform attracts a large enough pool of participants so that most points of views are represented). This should help mitigating information cascades by reducing the impact of the opinions of the people in the group (although it would not help mitigating reputational cascades). This should thus favour a debate, which should also improve the quality of arguments exchanged (as predicted by the argumentative theory of reasoning).

In support of this conclusion, Manin (2004) notes that “if the presence of contrary arguments [...] should be actively promoted so that a satisfactory deliberation can take place, it is relatively indifferent whether those arguments are first introduced without dialog”², and

² Translation from French: “*Mais si la présence d’arguments contraires [...] doit être activement favorisée pour qu’ait lieu une délibération satisfaisante, il est relativement indifférent que ces arguments soient d’abord introduits parmi les délibérants sans dialogue.*”

Mercier *et al.* (2017) argue that, to improve solitary reasoning, people should be “exposed to many [arguments that challenge their views]” because “once one has been exposed to counterarguments, it becomes much easier to anticipate them.” Though both focus specifically on opposing arguments (and not on diverse arguments), and neither talks specifically of consultative platforms, their common idea is that even if arguments come from a source which is external to the deliberating group or individual (here, a platform), it should still help them enrich their reflection.

One very important condition for this conclusion to be valid is that people evaluate arguments collected on platforms objectively. Indeed, for consultative platforms to be useful, decision-makers need to evaluate the arguments given by citizens relatively objectively, without just dismissing them because they come from an unknown source (a mere citizen) or because they support a divergent opinion. As noted by Mercier *et al.* (2017), “there is substantial evidence that people are good at evaluating arguments, at least when they care about the arguments’ conclusion”. Though some debate and uncertainty still exist as to the exact conditions under which people can evaluate arguments objectively (Trouche, Shao and Mercier, 2019), we make here the assumption that in at least some contexts, governmental decision-makers can keep their mind open to new arguments and counterarguments from citizens, whether because they have no strong preconceived opinion, or because, through mental exertion or some type of behavioural intervention, they become able to “tame” their cognitive biases to some extent.

Thus, existing literature provides evidence that collecting diverse arguments through consultative platforms may help decision-makers. Moreover, this conclusion is actually supported by some professionals of the sector. We got a confirmation of this during a semi-structured interview performed in March 2020, of a project manager at Cap Collectif – a firm which designs consultative platforms (the interview was performed face-to-face, recorded, and transcribed *verbatim*). Cap Collectif’s consultative platforms allow citizens to post arguments concerning different propositions, generally by distinguishing two columns: one for arguments in favour of the proposal and one for arguments against it. An example is given below (figure 2). The project manager explained that the goal of the consultative platforms was “to collect the largest possible diversity of opinions on a subject”³. “Opinions” here did not merely mean “unjustified preferences” but also arguments, as it became clear later in the interview. Indeed, when the project manager was asked about the criteria for a “successful platform”, she gave the following criterion (among others): “it is successful if the decision-maker has collected the **diversity of arguments**, and that it enabled him to modify his decision, to amend his decision, to explain his arbitrations”⁴. Thus, for Cap Collectif’s consultative platforms at least, an explicit purpose of the consultation is to collect the highest possible diversity of arguments, because it is believed that a diverse pool of arguments will allow decision-makers to improve their decision. Here, the interviewee did not make any distinction between the two dimensions of

³ Translation from French: “recueillir la plus grande diversité des opinions sur un sujet”

⁴ Translation from French: “[L’opération] est réussie en premier lieu si le décideur a récupéré la diversité des arguments, et que ça lui a permis de modifier sa décision, d’amender sa décision, d’expliquer ses arbitrages.”

diversity (non-redundancy and diversity of aspects) but, once again, it seems clear that having both would be the best option. We can hardly see why a decision-maker would ask to produce arguments about a question if they did not want to collect arguments that tackle all the aspects of the question (if not, they would probably have asked to produce arguments only on the particular aspect they were interested in).

Of course, a single example cannot prove that collecting diverse arguments is an explicit goal of many consultative platforms. However, it does show that some professionals of the sector themselves consider that collecting a diverse pool of arguments is an advantage, which can be seen as further evidence in support of our conclusion. Moreover, we would like to emphasize that this example is far from anecdotal, as Cap Collectif has designed an important number of consultative platforms for the national government and local governments in France, including very high profile consultations such as the *Grand Débat* consultation and the consultation about the universal income reform in 2019 (see here: <https://cap-collectif.com/realisations-2/>, for a list of all their platforms). This suggests that the goal of collecting diverse arguments is not just a whim of one civic tech firm but is shared by many public officials.

Thus, collecting diverse arguments appears as a legitimate goal of consultative platforms, because it is very likely to help making better decisions. Of course, not all decision-making groups are subject to cognitive similarity, laziness, bias and groupthink in the same degree, and many governmental decision-making groups take excellent decisions all the time, without needing to collect more diverse arguments through consultative platforms. Our point is not to say that every governmental decision-making groups should launch a consultative platform for every decision. It is only to show that *some* decisions could indeed benefit from collecting diverse arguments on consultative platforms.

Figure 2. A proposal in the platform “Vers un revenu universel d’activité” (“Towards a universal income for activity”) launched by the French government in 2019 (<https://www.consultation-rua.gouv.fr/>)

Proposition
1. Offrir un système plus lisible

Ministère des Solidarités et de la Santé • 7 oct. 2019 • Modifiée Épinglé
Regrouper et harmoniser un maximum d'aides sociales
3 538 votes • 1 097 arguments • 7 sources

Aujourd'hui, selon sa situation, une personne peut demander plusieurs aides différentes en même temps. Par exemple le **RSA** si elle n'a pas d'activité et les **APL** pour financer son logement.

Le **Revenu universel d'activité** vise à regrouper les **minima sociaux**, la **prime d'activité** et les **aides au logement** pour lutter plus efficacement contre la pauvreté.

D'accord Mitigé Pas d'accord

3 538 votes

Signaler Partager

1 097 arguments 7 sources

Ajouter un argument pour Ajouter un argument contre

628 arguments pour Les plus récents ▼

469 arguments contre Les plus récents ▼

Arguments proposed by previous participants

2. Research question and hypotheses: the effect of the exposure to other participants' arguments on aspect diversity

2.1. Research question

We have shown in the previous section that increasing argument diversity on platforms appears as a valuable goal. Both dimensions of diversity, i.e. the number of non-redundant arguments and the number of aspects tackled, appear *a priori* equally useful. It would thus be very useful to know how we can increase argument diversity, whether it be argument non-redundancy or aspect diversity, on consultative platforms. There are mainly two ways to do so. The first way is simply to attract more participants. The second way is to design the platform so as to collect the highest possible diversity of arguments from a fixed number of participants. How to incite citizens to participate in online participatory platforms has been extensively studied (e.g. Towne and Herbsleb, 2012; R. Farina *et al.*, 2013). On the other hand, to our knowledge, the second way, i.e. favouring argument diversity through design, has not yet been explored (this is not surprising since the importance of argument diversity on consultative platforms has not yet been recognized in existing literature). This is why we propose in this thesis to start exploring this second way.

We decided to focus on the effect of the exposure to other participants' arguments on aspect diversity. Thus, our research question is the following: **What is the impact of seeing other participants' arguments on aspect diversity on consultative platforms?** There are mainly two reasons for focusing on the exposure to others' arguments. The first reason is that existing literature allows to think that exposure to other participants' arguments is very likely to have an impact on aspect diversity. Especially, literature about idea brainstorming has shown that the exposure to other people's ideas has a significant impact on this dimension of idea diversity (e.g. Ziegler, Diehl and Zijlstra, 2000). It appears thus useful to question and test to what extent similar results can be found for arguments in "argument brainstorming" tasks, given the specific cognitive processes underlying argument production. The second reason is that platform designs largely vary on this point. At one end of the spectrum, there is for instance the Scottish Consultation Hub (see Figure 1), where participants do not have the opportunity to read other participants' contributions. At the other end of the spectrum, there is for instance the platform "*Vers un revenu universel d'activité*" (see Figure 2), where participants' arguments are immediately visible, without even needing to actively look for them or to go on another webpage. Thus, it is possible for platforms to choose different designs regarding the visibility of participants' arguments and they indeed do adopt different choices. This suggests that the visibility of participants' arguments is a debated design choice among platform designers.

In the following section we expose the specific hypotheses concerning the impact of the exposure to other participants' arguments that can be drawn from existing literature.

2.2. Hypotheses

2.2.1. *Main hypothesis: When other participants' arguments are visible, aspect diversity decreases at the group level*

In idea brainstorming tasks, studies have shown that seeing other participants' ideas decreases the number of semantic categories (which are equivalent to what we call aspects) tackled by groups, because members of the group focus on categories explored by previous participants instead of looking for new ones (see for instance Ziegler, Diehl and Zijlstra, 2000; for a review see Nijstad, Bechtoldt and Choi, 2019). Thus, seeing other participants' arguments could induce people to **focus on the aspects already mentioned** by previous participants, which would lead to a lower level of aspect diversity within the argument pool produced by the group.

The decrease in aspect diversity could moreover be strengthened by the will of people to **react to the arguments they were shown**: indeed, the *argumentative theory of reasoning* (Mercier, 2016; Mercier and Sperber, 2011) predicts that when producing arguments, people do not aim to discover the truth but to convince others. Seeing arguments against one's views might thus entice people to respond to those arguments (and so to deal with the same aspects) rather than trying to find new aspects that have not yet been tackled. Indeed, they will probably have more impact on their opponents' opinion if they devise precise counterarguments to their claims. Though in participatory democracy there is no argument exchange and so no hope of convincing one specific person, such an argumentative strategy could still happen if people consider the arguments they have seen as representative of those of the opposing group in general.

However, there is some evidence that in certain contexts people being confronted with opposing comments are less likely to talk about the same "topic" than people being confronted with comments they agree with (McInnis *et al.*, 2018). "Topics" in McInnis *et al.* (2018) are globally equivalent to what we call here aspects, that is to say large semantic categories that all relate to one general issue. The results of McInnis *et al.* suggest that participants might actually be more likely to tackle the same aspects as those evoked by arguments consistent with their views than to tackle the same aspects as those evoked by counterarguments. Nonetheless, this would *a priori* still lead to a decrease in aspect diversity.

Moreover, if people notice that the arguments they have read tend to focus on one aspect of the issue, they might choose to add their own thoughts on this aspect because of either **informational influence** or **normative social influence**. Informational influence would happen if seeing the arguments of others changed people's opinion on the importance and validity of certain aspects ("if people focus on this aspect, it must be because it is the most

important one”). It has already been shown that others’ opinions can have such an influence on people (Sunstein and Hastie, 2015), and induce them to ignore their private doubts about the validity of an opinion. Seeing other people focus on certain aspects could have a similar influence, and incite people to change their minds not about the best option but about the importance of those aspects.

Normative social influence would happen if, without changing their own opinion, people adapted their argumentative strategy for reputational purposes (“to have more credibility, I should focus on the aspects other people focus on”). There is already some evidence of the fact that focusing on *shared information* does have social benefits, through the “**mutual enhancement**” effect. Mutual enhancement refers to the fact that members of deliberating groups evaluate themselves and, more importantly, others as “more competent, knowledgeable, and credible” when they discuss shared information - information that all members already know before the discussion - than when they discuss unshared information – - information that only one member knows before the exchange (Wittenbaum and Park, 2001). The existence of a mutual enhancement effect has been shown by Wittenbaum, Hubbell and Zuckerman (1999) and proposed as an explanation for the bias towards shared information in deliberating groups. The shared information bias consists in group members sharing more and *repeating more* the information shared by all members of the group. Similarly, when facing arguments, people could be tempted to favour shared arguments or arguments that are based on shared information, thus tackling the same aspects as other participants. Indeed, it could induce a similar mutual enhancement effect. Of course, mutual enhancement is less likely to occur in an online anonymous context without group cohesiveness. However, some evidence exists that shared information bias still exists in computer-mediated groups (Lu, Yuan and McLeod, 2012) so it is possible that mutual enhancement also has some effect in online contexts. Thus, informational influence and normative social influence could lead to a decrease in the level of aspect diversity achieved by groups.

Finally, if other participants’ arguments are rather homogenous, seeing them may induce a **downward social comparison** (in a way, the reverse of social stimulation), which would result in a decrease in aspect diversity. In other words, if the other people tackle each only a few aspects and if these aspects are similar, participants may feel less inclined to make efforts to find numerous new aspects. However, evidence about the existence of a downward performance matching in brainstorming groups are mixed (see for instance Paulus and Dzindolet, 1993; Leggett Dugosh and Paulus, 2005).

2.2.2. Alternative hypothesis: When other participants’ arguments are visible, aspect diversity increases at the group level

Seeing other participants’ arguments could induce cognitive stimulation and/or social stimulation. Both effects have been shown by several studies to play a role in idea brainstorming tasks (e.g. Yagolkovski, 2016; Fink *et al.*, 2012; Leggett Dugosh and Paulus, 2005). **Cognitive**

stimulation occurs when seeing other participants' ideas "stimulate[s] group members to generate ideas that they would otherwise not have produced" (Dennis and Valacich, 1999). **Social stimulation** occurs when "an individual regards another person's rare idea as a high external standard, and an upward social comparison effect takes place [...], such that individuals' motivation to perform better is activated and they try to attain a higher level of creative activity to achieve a higher standard." (Yagolkovski, 2016).

Both cognitive and social stimulation have been shown to lead to an increase in the number of non-redundant ideas produced by participants in brainstorming tasks (see studies mentioned above), however, they have never been shown to lead to an increase in the number of aspects tackled. Thus, even though either of those phenomena could *theoretically* also lead to an increase in aspect diversity (if cognitive stimulation led people to think of new aspects or if social stimulation led people to increase their efforts to find new aspects), this outcome is unlikely.

3. Materials and methods

To test our hypotheses, we conducted an online experiment, whereby we asked participants to produce arguments about an imaginary public policy project (namely, the introduction of an urban toll in their city), as if they were on a consultative platform. In one condition (*condition A*), the participants did not see any arguments, in the other condition (*condition B*), they saw some of the arguments given by the participants of condition A. The details of the experimental design and methodology are exposed below.

3.1. Participants

3.1.1. Recruitment

224 participants were included in total: 25 participants for the pilot study, 100 participants for condition A, and 99 participants for condition B. They were recruited from the Prolific.co web-based population (www.prolific.co). We first recruited participants for condition A, then participants for condition B twenty days later. For each condition, we started by recruiting 100 participants, we then excluded those meeting one of the pre-registered exclusion criteria (see below in section 3.3.2. for the details of the exclusion criteria) and recruited new ones to replace them. We reiterated this procedure until we had 100 admissible participants for each condition. This follows the methodology we have detailed in our pre-registration (to see our pre-registration, use the following link : https://osf.io/qb8ve/?view_only=f70ef9fa2f0445e18f40a2b6ab1c0fb1). The excluded participants were removed without looking at the arguments they had produced or any other

data beyond what was necessary to determine their admissibility. In total 8 participants met one of the exclusion criteria for Condition A, and 14 for Condition B. 4 other participants also had to be excluded in Condition B because they noted in the “remark” section of the experiment that, due to a technical problem, they had not been able to read arguments of other participants.

For condition B, however, we had to exclude one more participant later on, for whom some data had not been recorded properly. We could not then recruit another participant, as the coding had already been completed. This is why data on condition B is based only on a sample of 99 participants.

3.1.2. Sample size rationale

Sample sizes in experiments dealing with electronic group brainstorming (whose methodology is the closest to the present experiment as they aim to evaluate the diversity of ideas produced by groups) go usually from less than 30 participants (as in Fink *et al.*, 2012) to about 150 participants (as in Yagolkovski, 2016), dispatched into the different test and control conditions, so that each group contains between 30 and 40 participants maximum. However, given the fact that we recruited participants from the internet, and that the task was rather short (less than 15 minutes) and completed online, it was possible to recruit a larger number of participants, and thereby to increase our statistical power.

3.1.3. Inclusion criteria concerning participants' profile

Participants had to be at least 18 years old. They had to speak French, as the task was to be performed in French. Moreover, they needed to be of French nationality and to live in France, so as to constitute a relatively homogenous sample in terms of culture, presence of urban tolls and knowledge /opinion about urban tolls. Indeed, research about urban tolls have shown some differences among countries in the opinion about such pricing schemes (see for instance Eliasson, 2016). Those different inclusion criteria were implemented through the recruitment filters proposed by Prolific.co.

3.2. Procedure

Participants completed the experiment entirely online. First, they had to perform the argumentative task (i.e. to produce arguments about the policy project). For this task, participants were assigned to one of two conditions. After that, participants answered a series of attention check questions to make sure they had well executed the task. Finally, they had to answer some questions regarding their individual characteristics.

3.2.1. The two conditions

Participants were assigned to one of two conditions:

- In condition A, no arguments were shown.
- In condition B, four arguments were shown.

The arguments shown to B-participants (i.e. participants from condition B) were selected among those produced in condition A. In order to ensure that the overall results were not dependent on a specific set of arguments, each participant saw a different set of arguments, chosen pseudo-randomly within the argument pool. The criteria used to choose the arguments were the following:

1. Each pool consisted of four arguments, with one argument for every non-neutral position (against, rather against, rather in favour, in favour). Arguments coded as being “neither against nor in favour” were excluded to ensure that participants were not exposed to confusing arguments. The arguments were equally distributed over the four non-neutral positions to avoid any social pressure in favour of one position, as it might trigger some cognitive processes that could interfere with our results. B-participants were explicitly told the position of each argument they saw, so that no ambiguity was possible.
2. Each pool of 4 arguments had to tackle at least two aspects. This was done to guarantee that all participants were exposed to a minimum aspect diversity. Indeed, a total absence of aspect diversity is likely to have specific effects, which are not to be tested in this particular experiment.
3. Participants were not exposed to arguments that were coded as tackling no particular aspects, so that they were not exposed to confusing or very poor arguments. Indeed, exposure to arguments of very poor quality could have specific effects, which are not to be tested in this particular experiment.
4. Finally, participants were not exposed to posts that had been coded as containing several arguments. This was done to ensure that all participants saw strictly one argument per position.

3.2.2. The argumentative task

Why an urban toll?

The fictional public policy project consisted in the possible introduction of an urban toll around the city centre of the city they lived in or visited the most regularly. There are four reasons why urban tolls were considered as an appropriate subject. Firstly, studies in France have shown urban tolls to be a controversial issue (e.g. Souche-Le Corvec *et al.*, 2016). Using such a controversial issue insured collecting a large number of arguments for both sides of the

debate, which was essential to ensure that B-participants could read arguments both for and against the policy. Moreover, it was important for the ecological validity of our study that our fictional policy proposal could be the object of a real debate: *a priori*, consultations are not launched for policy proposals which are totally consensual.

Secondly, urban tolls are *complex* public policies. They have multiple aspects, and multiple arguments can be found for and against them. A simple proof of that is the number of studies and reports about urban tolls (e.g. ADEME, 2014), their economic consequences (e.g. Kopp and Prud'homme, 2010), their acceptability (e.g. Raux and Souche, 2004), etc.

Thirdly, as yet no urban toll has been introduced in France, and there has been no major nation-wide discussion on urban tolls since 2018. This ensured that people were unlikely to have recently been exposed to arguments about this issue. This is important as we want to measure the impact of exposure to arguments from other people, so an exposure to a large pool of arguments outside of our experiment would bias our results.

Finally, almost all citizens are concerned by the issue, as it tackles pollution, environment, social justice and car use. Thus, participants were likely to have an interest in the issue and an earnest opinion about it, and as a consequence were likely to be reasonably motivated to produce relevant arguments reflecting their opinion. This should help avoid really poor arguments, or arguments produced “at random” without any real reflection on the subject. This allowed us to be as close as possible to a real consultation context, where only people who have an interest in the issue self-select to participate.

The instructions

Participants were first presented with a description of urban tolls. This aimed to ensure that all participants knew the main characteristics of this particular type of road pricing scheme. The description was as neutral as possible, and only evoked the main characteristics of urban tolls, so that participants would know enough to understand this public policy but would not be biased or inspired to produce arguments on particular aspects. Participants were then presented with a description of the fictional project of urban toll they would have to comment on, and asked to imagine they were on a real consultative platform. Once again, the description was as neutral as possible, to avoid bias.

Participants were then presented with the instructions they had to follow when producing arguments. The instructions were meant to clarify the type of contributions participants were expected to write. In particular, a definition of “argument” was given (see Appendix I, which includes all the instructions presented to participants).

Participants from both conditions were told that their arguments could be shown to future participants. Three reasons justify this decision. First, for ethical reasons, as we indeed showed the arguments of A-participants to other participants, we needed to be transparent with A-participants on this point. Second, to test only the effect of exposure to others' arguments

ceteris paribus, we needed both conditions to differ as little as possible, so we had to tell the same thing to both series of participants. Finally, in real life consultative platforms, whatever the design, arguments are supposed to be read at least by the decision-makers and to be visible to all citizens in a final public report (although they will be more or less summarized in that report). Thus, in real life, arguments are always meant at some point to become public (at least to some extent), and to have a chance of changing some people's minds. We feared that if participants had no reason to believe that their arguments would have a chance of influencing someone's opinion, they might lack the motivation to produce arguments, *a fortiori* good arguments. Thus, we believe that the fact of telling the participants that their arguments could be shown improves the ecological validity of our results.

For condition B, specific instructions were added concerning the arguments shown to participants. B-participants were told that the arguments they saw were chosen at random. This was meant to prevent participants from believing that the experimenters had carefully selected the arguments they saw for some hidden experimental purposes, which might have biased them by inciting them to try to guess the purpose of the experiment based on the features of the arguments they were shown. Moreover, we explicitly told participants that some of the previous participants had produced several arguments to prevent false beliefs about the number of arguments written by others from creating a bias. Indeed, as only one argument per position was shown, B-participants could have thought that it meant that most of the previous participants had produced only one argument each. This could have biased B-participants towards following this apparent social norm and producing only one argument. Though a mere statement may not be enough to totally exclude the biasing effect of the apparent social norm, we believe it can reduce it significantly.

The organization of the argumentative task

After having been shown the different instructions, participants could produce as many arguments as they wished (as in any real-life consultative platform). They were given the possibility to access the description of urban tolls and the instructions at every moment while completing the task. In condition B, they could also have access to the arguments they had read. There was no time limit for either reading or writing arguments. However, participants were compelled to spend a certain amount of time reading the presentation of urban tolls (30 seconds) and the description of the urban toll project (35 seconds), to ensure they had read them thoroughly. A minimum length of 40 characters was required for each argument they posted. After having produced all the arguments they wanted, they were required to select a position for each argument they had posted (between "in favor", "rather in favor", "neither against nor in favor", "rather against", and "against").

3.2.3. Attention check questions

Participants were required to answer five multiple choice questions about urban tolls in general and about the specific project of urban toll to make sure that they had understood this public policy and that they had thoroughly read the instructions. The questions are in the appendix. Following the preregistration, the participants who answered wrongly to two questions or more were excluded. Besides, B-participants were asked a series of five questions about the arguments they had read. First, they were asked whether there was a majority of arguments for or against urban tolls, or if the two opinions were evenly represented. Then, they were presented with 4 pairs of arguments. Each pair of arguments contained one argument they had read and one other argument of the same position (in favour, rather in favour, etc.) chosen randomly. They were asked to identify for each pair which one of the two arguments they had read before. Following the preregistration, participants had to answer wrongly to three questions or more in this series of five to be excluded.

3.2.4. Questions about individual characteristics

Participants were first asked a series of question about their car use and their interest in environmental questions. Car dependency and environmental concerns have indeed been shown by several studies (e.g. Jaensirisak, Wardman and May, 2005; Eliasson and Jonsson, 2011; Souche-Le Corvec *et al.*, 2016) to have an important influence on people's opinions about urban tolls. Thus, it was important for us to see if there were large differences in levels of car dependency or environmental concerns between A-participants and B-participants, because it might have induced participants to tackle different aspects independently of our experimental design. For car use, we used the same question as Souche-Le Corvec *et al.* (2016) and for environmental concerns, we performed a translation (as faithful as possible) of the questions used by Eliasson and Jonsson (2011).

Then, participants were asked if, during the week preceding the experiment, they had participated in or witnessed a discussion about urban tolls. This aimed to check that they had not already had the opportunity to reflect about diverse arguments outside of the experiment. Participants were also asked if they knew before the experiment what an urban toll was (No – More or less – Yes). This question aimed to ensure that observed differences between condition A and condition B were not due to an important difference in previous knowledge about urban tolls. Indeed, having some previous knowledge about urban tolls might make it easier for participants to think of a large number of aspects and/or of less common aspects.

Finally, participants were asked a series of demographic questions: their gender, their age, their level of education, their profession and their political affiliation.

3.3. Data analysis

3.3.1. Data exclusion

Following the preregistration, we used 3 exclusion criteria concerning participants. They were excluded:

1. If they had participated in or witnessed a discussion on urban tolls in the last week.
2. If they spent much more or much less time than others completing the task (± 3 standard deviations from the mean).
3. If they had failed to answer correctly the attention check questions.

Moreover, arguments were excluded when they failed a minimum relevance check. The relevance check included the following criteria:

1. Arguments must deal with urban tolls.
2. Arguments must justify the participant's opinion. This leads to exclude for instance mere factual declarations.
3. Arguments must be based on a characteristic of urban tolls and not on a pure idiosyncrasy of the participant (for instance: "I am against urban tolls because it would increase my car budget" is based on the fact that urban tolls imply paying a fee, so it is acceptable. On the contrary, the argument "I am against urban tolls because I am in a bad mood" is not based on any objective characteristic of urban tolls, so it would be excluded).

Initially, a "complete sentence" criterion (i.e. the arguments had to contain at least one complete sentence) was also included. However, the pilot study showed that even arguments that did not contain a complete sentence were understandable and relevant, so this criterion was abandoned.

3.3.2. Dependent variables

Main dependent variable

As our hypotheses are formulated at the group-level, our main dependent variable is a group-level variable:

- **Variable 1: The probability that groups of n participants will tackle at least x aspects.**

Variable 1 aims at comparing the level of aspect diversity achieved by same-sized groups from the two conditions. Importantly, for B-groups (i.e. groups from condition B), we included in the counting the aspects tackled by the 4 arguments presented to each participant. For A-groups, we also assigned once and for all to each participant 4 arguments chosen pseudo-randomly from the pool of arguments produced by A-participants, and included in the counting

the aspects tackled by those arguments (we made sure that A-participants were not assigned their own arguments). As the pools of assigned arguments were both selected from condition A using the same criteria, the pools assigned to A-participants and B-participants should be on average equivalent, as far as aspects are concerned. Thus, the only difference between A-groups and B-groups is that A-participants have not seen the arguments that have been assigned to them whereas B-participants have. Thus, comparing same-sized groups for both conditions allows us to answer the following question: does the fact that members of B-groups have seen their assigned arguments impact the level of aspect diversity achieved by the group?

In order to test a large range of group-sizes with our samples (going up to groups of 90 participants), we used a resampling technique, whereby for each group-size (and for each condition), we constituted randomly 1000 groups of said size. We controlled that no two groups were identical. For a group i , we will note y_i the number of aspects tackled by that group. In each condition, for each group-size n , we measured the probability that y_i would be equal or superior to a certain number x . In other words, we measured the probability that a group of n participants would tackle at least x aspects. We will note this probability for condition A, $P_{n,A}(y_i \geq x)$ and for condition B, $P_{n,B}(y_i \geq x)$. We calculated this probability in both conditions for groups of 2 to 5 participants and for groups of 10, 20, etc., up to 90 participants. We also calculated it for all possible numbers of aspect (x) going from 1 to 11.

Complementary individual-level variables

We also introduced some complementary variables to collect data on the effect of exposure to others' arguments at an individual level. Variables 2 and 3 aim to better understand how group-level effects possibly detected by Variable 1 might be explained by individual-level effects. Variables 4 and 5 give general information about how exposure to others' arguments impacted individual behaviour relative to the number of arguments produced and the number of aspects tackled.

➤ Variable 2: The level of aspect originality per participant.

Variable 2 corresponds to the number of original aspects tackled per participant. "Original aspects" are aspects tackled by the participant that were not present in the arguments assigned to him/her. Thus, this variable measures the propensity of a participant to find aspects that were not already tackled by the four arguments assigned to him/her. It is important to note on this point that the tackling of an original aspect is not determined by the same process for A-participants and B-participants. On the one hand, B-participants have read their assigned arguments, and are thus aware of which aspects have been tackled by their assigned arguments. Thus, in their case, tackling an original aspect can be due to a voluntary effort to search for aspects that were not already tackled. On the other hand, A-participants did not read the arguments assigned to them. Thus, in their case, finding original aspects can of course *not* be due to a voluntary search for new aspects. Thus, the number of original aspects tackled by A-participants is only determined by how common (i.e. how frequently tackled in the general sample) the aspects they tackle are.

This variable gives us valuable information about whether group-level effects could be explained by a difference in the number of original aspects tackled at the individual level. More specifically, it allows us to answer the following question: if the aspect diversity of B-groups is lower (higher) than for A-groups, can it be due to the fact that B-participants individually tackle less (more) original aspects than A-participants? In other words, can it be due to the fact that B-participants achieve a lesser (higher) level of aspect originality?

- **Variable 3: The number of aspects *all arguments included* per participants (i.e. when both assigned arguments and produced arguments are taken into account for each participant).**

In both conditions, being assigned more aspects makes it mechanically more difficult to find numerous original aspects. Thus, taking into account the number of assigned aspects at the individual level allows to take into account this inequality between participants depending on the number of aspects they have been assigned. This is why it is important to introduce Variable3, as a complement to Variable 2, in order to better evaluate individual effects.

- **Variable 4: The number of arguments produced per participant.**

Variable 4 allows us to see whether exposure to other participants' arguments has an impact on the number of arguments participants write or only on the content (aspect-wise) of their arguments.

- **Variable 5: The number of aspects tackled per participant.**

Variable 5 allows us to establish whether exposure to other participants' arguments has an impact on the number of aspects tackled per participant, be they original or not.

3.3.3. Coding methodology

The coding steps

Coders had to perform 4 coding steps for each post:

1. Coders determined whether the post was relevant (i.e. passed the relevance check).
2. They determined whether the post contained one or several arguments.
3. They coded the position of each argument contained by the post, correcting, if need be, the position indicated by the participant.
4. They coded the aspect(s) tackled by each argument. Importantly, one argument could tackle several aspects. Indeed, it is not unusual for arguments to establish links between several different elements which fall into different aspects.

Development of the coding scheme

An original coding scheme was developed, and corresponding coding instructions were written and given to all coders involved. As underlined by Neuendorf (2017), a coding scheme should be revised several times during coder training so that the experimenters and the coders are all comfortable with the coding scheme and that there are no ambiguities left. However, coders could not be trained before the pilot, which limited the opportunities of revision of the coding scheme. Nonetheless, the coding of the pilot enabled an experimenter to discuss with coders to identify existing ambiguities in the coding guidelines. Thus, the coding scheme and the corresponding instructions were revised twice, after the first coding and the second coding of the pilot. Those two revisions resulted in a list of 12 aspects, plus one category “no aspect tackled”, and one category “other aspect”. The list of aspects is presented in the appendix.

Coding procedure

One experimenter and three lay coders (let’s call them coders 1, 2 and 3) were involved in the coding process. The table below details the role of each one.

Table 1. Coders involved at each stage of the coding process

Sample coded	Coders involved
Pilot (first coding)	Experimenter and Coder 1
Pilot (second coding on aspects)	Coder 2 and Coder 3
Condition A	Coder 1 and Experimenter (<i>subsample</i>)
Condition A+B	Coder 2 and Experimenter (<i>subsample</i>)

A first coding of the pilot study was performed by Coder 1 and the experimenter. The coding scheme was revised, and a second coding of aspects was performed by Coders 2 and 3, to make sure that the revisions had allowed to solve the identified problems.

Then, a coding including only A-participants was performed. This was necessary because coding arguments from condition A was a prerequisite for creating the argument pools shown to B-participants. The experimenter recoded a subsample of 50 posts (22% of the sample) to measure the coding reliability. The size of the subsample was determined following the guidelines proposed by Neuendorf (2017), i.e. it represented more than 10% of the full sample and contained no fewer than 50 units of data. The subsample was constituted randomly.

Finally, a final coding was performed by Coder 2 on the arguments from condition A and B. It was indeed important for our measures that all the arguments were coded by the same coder, so that the coding of both conditions would be coherent and based exactly on the same arbitrations. The posts were randomized, so that Coder 2 could not know from what condition

came the arguments he coded. Here too, the experimenter recoded a subsample of arguments (115 posts, which corresponds to 28% of the total, including 52 posts from condition A and 63 posts from condition B) to evaluate coding reliability. Once again, the subsample was constituted randomly, and the experimenter did not know either from what condition each argument came.

Considerations regarding coders, coder training and the coding method

Coders all spoke French fluently. They did not have any experience with similar coding schemes. They were not familiar with urban tolls. They were required to adopt the position of an *intelligent but ordinary external reader*.

Coders 1 and 2 were familiar with the hypotheses of the experiment. This is a limit of our coding methodology, as knowing what the experimenter hopes or expects creates an important risk of biasing the coding. However, coders were never put in a situation where they could compare arguments from condition A with arguments from condition B. This limits the possibility of a significant bias. Coder 1, when coding condition A, knew what condition he was coding, but could not compare the arguments he had with arguments from the other condition. Moreover, and most importantly, Coder 2 and the experimenter, when they performed the final coding of both conditions, were blind to the specific condition the arguments they were coding came from. Thus, the final coding at least could not have been biased.

Coders, who were volunteers, could not be trained before the pilot. This is a problem, because, as emphasized by Neuendorf (2017), training is an important step to ensure that the coding will be reliable and efficient. Nonetheless, every coder who was involved in the final coding had the opportunity to code the pilot (at least for aspects) and discuss with an experimenter, which constitutes a form of minimal training. Still, the absence of extensive coder training and the fact that the coding scheme was only revised twice, based on a pilot of 25 participants created two problems:

1. The coders could still have some trouble understanding the more subtle points of the coding instructions, and some of the more complex cases were not treated in the instructions (because they had not been encountered during the pilot).
2. Untrained coders are more likely to make careless mistakes. This last danger was confirmed, as three mistakes concerning the coding of aspects were noticed in the subsample of condition A analysed by the experimenter. No mistake was spotted concerning the other coding steps, probably due to the fact that they were less complicated and less likely to lead to typos.

To address the first point, it was decided that coders would be allowed during the coding to ask the experimenter some questions about how to interpret coding instructions or about how

to handle ambiguities in complex cases. However, coders would **not** be allowed to quote the arguments. The experimenter was allowed to make the rules clearer and more precise, but not to tell how to code specific arguments. Most of the time, the experimenter only repeated or explained the coding instructions, providing (fictional) examples or rephrasing some points to make them clearer. In some cases, the experimenter, judging the coding instruction not precise enough, added further guidelines. All these cases were noted and a written summary of the new guidelines was given to coders regularly so that they could check they had followed them.

To address the second point and avoid mistakes for aspects, it was decided that coders would follow the following proofreading methodology: after having performed a first thorough coding of the aspects, they would hide this first coding, and perform a second one, going quickly through the arguments. Then, they would compare the first thorough coding with the second quick one. As there is little chance that they would make twice the same careless mistake on the same argument, this methodology was likely to allow coders to detect most of the mistakes, if not all.

4. Results

4.1. Pilot study

We conducted a pilot study with 25 participants (who did not see any argument). This pilot study aimed at checking several points:

1. The overall duration of the experiment
2. The number of participants that would be excluded according to our exclusion criteria. Indeed, if an important number of participants were excluded, it would be necessary to change either the criteria or the design of the experiment.
3. The reliability of our coding scheme. Two coders performed a first coding, and a second pair of coders then recoded the aspects (see below in the “aspects” part). Especially, we wanted to see if participants could classify their own arguments correctly or almost correctly as far as aspects were concerned, which would have facilitated the coding afterwards.

4.1.1. *Exclusion criteria*

Four pre-defined participants were excluded either because they failed to answer the check questions or because of the time they spent on the task. Two arguments were excluded because they failed the minimum relevance check (coders initially disagreed on one of the two and solved this disagreement through discussion). Four arguments were excluded from the first coding because they did not contain a complete sentence, but were included in the second coding because the “complete sentence” criterion was abandoned.

Thus, very few participants and arguments were excluded, which confirms that the urban toll was an adequate choice of subject and that our instructions were clear enough to enable people to understand the urban toll project and to produce relevant arguments about it. It also confirms that our exclusion criteria are coherent with participants' behaviour in our task.

4.1.2. Coding reliability

Posts' relevance, division of posts, and arguments' position

Arguments from participants who failed to answer the attention check questions were excluded. Arguments which did not contain a complete sentence were also excluded at this point. However, the arguments from the one participant that took too much time to do the task remained in the analysis: since the pilot aimed at evaluating the reliability of the coding scheme, the more relevant arguments were coded, the better. Thus, 49 posts were coded by Coder 1 and an experimenter.

Cohen's kappa was measured. For arguments' position, a kappa with linear weighting was also measured, because it is more representative of coders' agreement on this point : indeed, a disagreement should not have the same weight whether it is between "against" and "rather against" or between "in favour" and "against". A kappa of 0.6 was generally taken as a threshold for acceptable agreement following the guidelines of Neuendorf (2017). The most rigorous threshold would be a kappa of 0.8 but this was hard to obtain because, due to the small number of posts, one single disagreement could make a very large difference in the kappa. This is why we also consider the percentage of intercoder agreement. For arguments' position, posts which both coders agreed to divide were divided and categorized separately, the others were treated as one argument. The same was done for the coding of aspect.

Table 2. Measures of intercoder reliability for the coding of post's relevance, division of posts, and arguments' position in the pilot

Criterion	Intercoder agreement (%)	Cohen's kappa
Post's relevance	98	0.66
Division of posts into single arguments (out of 47 posts, the irrelevant arguments being excluded)	96.2	0.48
Arguments' position (out of 49 arguments)	87.8	0.84 (Kappa with linear weighting: 0.91)

The coding appears quite reliable. Indeed, Cohen's kappa is above 0.6 for post's relevance and arguments' position. Though it is quite low for the division of posts, this is due to the very small number of posts containing more than one argument. The very high percentage of intercoder agreement allows to think that the coding is nonetheless reliable. The experimenter discussed with Coder 1 and revised the coding scheme to remove the existing ambiguities, especially regarding which posts were to be divided into several arguments.

Aspects

First coding

The first coding was performed by one experimenter and Coder 1. One of them forgot to code aspects for one argument. Thus, intercoder reliability is measured on a pool of 48 arguments. As participants from the pilot had been asked to code their arguments themselves as far as aspects were concerned, coders had access to participants' coding and had to correct it.

Cohen's kappa was measured. It was measured for each category separately. Indeed, as one argument could be classified into several aspects, we assumed that the classification could be best modelled in the following way: for each argument, coders decided for each aspect if it was present or not. Thus, we measured Cohen's kappa for each aspect, in a two-category model. Here too, a kappa of 0.6 was taken as the threshold for acceptable agreement. As some aspects were rarely tackled, which meant a single disagreement could make a very large difference in the kappa, other measures of coding reliability were added to give a more complete picture of coders' agreement: the total number of agreements and disagreements and the mean number of agreements and disagreements per category. Importantly, for aspects, one disagreement corresponds to one case where one coder thought an aspect was present when the other thought it was not. Thus, for instance, if one coder put aspect A while the other put B, it is counted as two disagreements: one disagreement being about the presence of A, and the other being about the presence of B. One agreement corresponds to one case where both coders agreed that an aspect was present (the cases where both coders agreed that an aspect was not present are not taken into accounts).

The results are the following:

- There were 18 categories in total. Aspects were divided into effects of the urban toll on the one hand, and characteristics of the urban toll on the other hand.
- For 10 categories, Cohen's kappa was above 0.6, for the 8 other categories, the kappa was below 0.6. Thus, for 8 categories (almost half of our categories), the agreement between coders appeared very unsatisfying.
- In total, there were 83 agreements (4.6 agreements per category) and 89 disagreements (4.9 disagreements per category). The very high number of disagreements confirms that the coding was unreliable.

The first coding thus brought to light several issues regarding the methodology for coding aspects. First, several aspects were ambiguous. Moreover, participants did not code their own arguments properly: coders considered that less than half of the arguments had been properly coded by participants (Coder 1 agreed with participants' coding for only 23 arguments out of 48 and the experimenter agreed with participants' coding for only 16 arguments out of 48). We suspect that many participants based their coding on what they meant rather than on what they had actually written. Thus, it was decided that participants would not code their own arguments. Finally, the experimenter discussed with Coder 1 and identified the reasons for their disagreements. It appeared that dividing aspects between characteristics and effects was misleading, thus this division was abandoned. A new list of aspects, along with more precise coding instructions, were elaborated. They were tested with two different coders in a second coding.

Second coding

As it appeared that arguments that did not contain a complete sentence were actually still understandable and relevant, this exclusion criterion was abandoned and the arguments without a complete sentence were included in the second coding. Thus, the second coding was performed on 53 arguments, with the revised categories.

The results are the following:

- Including one category “no aspect mentioned”, and one category “other aspect”, there were 14 categories. This time, the category “other aspect” was not used, which suggests that our list of aspects was exhaustive.
- Cohen's kappa was measured for each category: only 3 categories were below 0.6. Thus, for most categories, the coding appeared quite reliable.
- In total, there was 53 agreements (3.8 per category) and 37 disagreements (2.6 per category). If we compare this to the first coding, we see that the total number of disagreements was divided by more than two and that the mean number of disagreements per category also showed a sharp decrease. This confirms that there was much less ambiguity in the revised coding scheme regarding when aspects were (or not) tackled. We can notice there are also less agreements in total and per category but this is due to the new coding guidelines which were stricter regarding when an aspect could be considered as present, thus leading to a general decrease in the frequency of aspects.

Thus, the second coding showed significant improvement in the reliability of the coding scheme. Coders discussed with the experimenter, and the reasons for the disagreements were identified. The list of aspect was not changed, but the coding instructions were adapted to solve the remaining ambiguities, by adding examples and more precise guidelines regarding the exact delimitations of aspects.

Based on the results of the pilot, we realised a pre-registration, which included especially the results of the pilot, the coding instructions, and the final list of the exclusion criteria.

4.2. Main study

4.2.1. Comparison between the two participant samples (A and B)

Table 3. Comparison between the two participant samples regarding demographic profile

	Condition A	Condition B	t-value	X- squared	df	p
Age	<u>28.59</u>	<u>27.36</u>	0.950	/	197	.343
Gender:			/	1.581	3	.664
Man	61.0	57.6				
Woman	37.0	40.4				
Other	2.0	1.0				
Does not wish to say	0.0	1.0				
Level of education:			/	5.971	5	.042*
<i>Brevet des collèges</i> (secondary school diploma)	1.0	0				
<i>CAP, BEP</i> , other same-level diploma	0.0	1.0				
<i>Baccalauréat (BAC) général, technologique,</i> <i>professionnel</i> or equivalent	14.0	19.2				
<i>Licence, BTS, DUT</i> , other <i>BAC+2</i> qualification	28.0	35.4				
Master's degree	54.0	35.4				
PhD	2.0	9.1				
Does not wish to say	1.0	0				
Profession:			/	1.970	6	.923
Craftsmen, tradesmen, shopkeepers, heads of business	2.0	3.0				
Managers and highly qualified professionals	28.0	25.3				
Intermediary professions	10.0	8.1				
Employees	17.0	16.2				
Factory workers	0.0	0				
Retired	1.0	0				
Other people without employment	33.0	38.4				
Does not wish to say	9.0	9.1				
What political party do you feel closest to?			/	13.221	8	.105
Le Front National	0.0	5.1				
Les Républicains	4.0	4.0				
En Marche	14.0	7.1				
Europe Ecologie Les Verts	8.0	11.1				
Le Parti Socialiste	12.0	10.1				
Les Insoumis	10.0	6.1				
Other	2.0	5.1				
None	35.0	43.4				
Does not wish to say	15.0	8.1				

* $p < .05$ (it is generally considered that a p-value below 0.05 indicates a statistically significant difference).

df = degrees of freedom

For age, the mean (underlined in the table) of each sample is presented, and p was measured with a t-test. For the other variables, the percentage of participants falling into each category is presented, and p was measured with a chi-squared test.

Table 4. Comparison between the two participant samples regarding toll-related traits

	Condition A	Condition B	t-test	X-squared	df	<i>p</i>
I am [...] in environmental issues (on a scale from 1 – “not at all interested” – to 5 – “very interested”)	<u>4.08</u>	<u>4.10</u>	-0.181	/	197	.857
How important is it for you that you travel in an environmentally friendly way? (On a scale from 1– “not at all important” – to 5 – “very important”)	<u>3.68</u>	<u>3.71</u>	-0.202	/	197	.840
Do you worry about environmental issues? (on a scale from 1 – “No, never” – to 4 – “Yes, often”)	<u>3.36</u>	<u>3.39</u>	-0.342	/	197	.733
Do you use a car:			/	1.173	3	.760
Never or rarely	41.0	44.4				
At least twice a month	19.0	14.1				
At least twice a week	20.0	23.2				
Every day	20.0	18.2				
Did you know what an urban toll was before this experiment?			/	2.19	2	.335
No	23.0	32.3				
More or less	38.0	34.3				
Yes	39.0	33.3				

For environmental concerns, the mean (underlined in the table) of each sample is presented, and *p* was measured with a t-test. For the other variables, the percentage of participants falling into each category is presented, and *p* was measured with a chi-squared test.

Participants from both conditions have very similar demographic profiles, except for education level: $X^2(5; N=199) = 5.971, p = .042$; all other *p*-s $>.104$ (see table 3 and table 4). More importantly, both samples are comparable in terms of toll-related traits: whether it be environmental concerns (interest in environmental issues, importance of travelling in an environmental-friendly way, worry about environmental issues: all *p*-s $>.732$, see Table 4), level of car use, $X^2(3; N=199) = 1.173, p = .760$, or previous knowledge about urban tolls, $X^2(2; N=199) = 2.19, p = .335$. As those three elements were especially likely to influence the aspects tackled by participants, such similarities confirm that the observed differences between the two conditions in the aspects tackled are robust.

4.2.2. Coding reliability

Coder 1 performed a first coding of arguments collected in condition A, and Coder 2 performed the final coding of both conditions. To measure intercoder reliability, an experimenter recoded a subsample of 50 posts for condition A, and a subsample of 115 posts

for the final coding. Results are detailed below. We adopted the same measures of coding reliability as in the pilot.

Condition A

Posts' relevance, division of posts, and arguments' position

Table 5. Measures of intercoder reliability for post's relevance, the division of posts, and arguments' position in condition A (on a sample of 50 posts)

Criterion	Intercoder agreement	Cohen's kappa
Posts' relevance	100%	1
Division of posts into single arguments (out of 49 posts, the irrelevant argument being excluded)	90 %	0.61
Arguments' position (out of 55 arguments)	91%	0.87 (Kappa with linear weighting: 0.93)

Precision regarding the division of posts : six posts were divided by both coders, but among those six posts, one was divided in two by one coder and in three by the other. This was counted as a disagreement.

The measures suggest that the coding is reliable. Indeed, the percentage of intercoder agreement is systematically equal to or above 90% and Cohen's Kappa is above 0.8 for arguments' relevance and arguments' position, and above 0.6 for the division of arguments.

Aspects

For the coding of aspects, three "careless mistakes" due to lack of attention were noticed. Two mistakes were made by Coder 1 and one by the experimenter. In one case they wrote one digit instead of another, in two others, they forgot to add one aspect although they had added it in similar arguments. In order to make sure that the coder's mistakes were really due to a lack of attention, the experimenter asked him to recode the arguments without looking at his previous coding, and then asked him to confirm that the discrepancy between the initial coding and this recoding was indeed due to a lack of attention. Reliability was evaluated after the mistakes had been corrected (results before correction are in appendix). To avoid such careless mistakes in the future, a proofreading methodology was elaborated in concertation with coders (see above in section 3.3.4.).

Table 6. Measures of intercoder reliability for the coding of aspects in condition A (on a sample of 55 arguments)

Aspect	Number of agreements	Number of disagreements	Cohen's kappa
0. No precise aspect	1	1	0.66
1. Decision process	1	0	1
2. Money collected	9	0	1
3. Resources necessary to set up the toll	0	1	0
4. Perimeter	0	0	1
5. Technologies used to pay the toll	1	0	1
6. Rate	6	1	0.91
7. Who pays the toll?	11	4	0.80
8. Travel behaviours	6	8	0.51
9. Travel conditions and infrastructures	17	3	0.88
10. Pollution, environment	13	0	1
11. Quality of life	7	2	0.85
12. Economic and social circumstances	4	3	0.70
13. Other aspect	0	0	1
Total	76	23	/
Mean (per aspect)	5.4	1.7	/

Here also, the coding appears reliable enough to guarantee quite robust results. For aspects, two categories have a kappa below 0.6 (categories 3 and 8). For category 3, we do not consider the kappa to be meaningful because the category appears too rarely (only once). For category 8, an ambiguity was noticed between categories 8 and 9. As a consequence, in order to make sure that participants in condition B were exposed to arguments containing at least two different aspects, we checked that arguments shown to participants in condition B did not contain *only* aspects 8 and 9.

Final coding (conditions A+B)

Posts' relevance, division of posts, and arguments' position

Table 7. Measures of intercoder reliability for post's relevance, the division of posts, and arguments' position in the final coding (on a sample of 115 posts)

Criterion	Intercoder agreement	Cohen's kappa
Posts' relevance	100%	1
Division of posts into single arguments (out of 115 posts, the irrelevant argument being excluded)	93 %	0.76
Arguments' position (out of 135 arguments)	96%	0.94 (Kappa with linear weighting: 0.97)

Precision regarding the division of posts : seventeen posts were divided by both coders, but among those seventeen posts, one was divided in two by one Coder 1nd in three by the other. This was counted as one disagreement. Besides, two arguments were divided along slightly different lines by both coders, but as this did not change the number of arguments, nor the aspects tackled by the post as a whole, such disagreement is not important for our analysis, thus they were not counted as disagreements. Coders solved those two slight disagreements through discussion.

The results suggest that intercoder reliability is high enough to guarantee robust results. Indeed, the percentage of intercoder agreement is systematically above 90% and Cohen's Kappa is above 0.9 for arguments' relevance and arguments' position and is above 0.7 for the division of arguments.

Aspects

The coding of aspects was performed on 135 arguments: just like for arguments' position, posts which both coders agreed to divide were divided and each argument was coded separately, while posts which both coders did not agree to divide were treated as one argument.

For the coding of aspects, two careless mistakes due to lack of attention were noticed: one by the experimenter, and one by the lay coder. In both cases, they had forgotten to put one aspect down even though they had put that aspect down in similar cases. In order to make sure that the coder's mistake was really due to inattention, the experimenter asked him to recode the argument without looking at his previous coding, and then asked him to confirm that the discrepancy between the initial coding and this recoding was indeed due to a lack of attention. Reliability was evaluated after those mistakes had been corrected (results before correction are in the appendix).

Table 8. Measures of intercoder reliability for the coding of aspects in the final coding (on a sample of 135 arguments)

Aspect	Number of agreements	Number of disagreements	Cohen's kappa
0. No precise aspect	0	0	1
1. Decision process	2	0	1
2. Money collected	4	2	0.79
3. Resources necessary to set up the toll	3	1	0.85
4. Perimeter	0	0	1
5. Technologies used to pay the toll	2	0	1
6. Rate	12	1	0.96
7. Who pays the toll?	37	5	0.91
8. Travel behaviours	23	16	0.67
9. Travel conditions and infrastructures	50	14	0.79
10. Pollution, environment	28	2	0.96
11. Quality of life	19	3	0.91
12. Economic and social circumstances	12	5	0.81
13. Other aspect	0	0	1
Total	192	49	/
Mean (per aspect)	13.7	3.5	/

Here too, intercoder reliability is high enough to guarantee robust results: all categories are above 0.6 and only one category is below 0.79 (category 8).

4.2.3. Statistics about the arguments collected

409 posts were collected, including 227 for condition A, and 182 for condition B. The measures below are based exclusively on the final coding performed by Coder 2. Two posts were excluded because they failed the pre-registered relevance check. As participants who produced those posts also produced other posts, this did not lead us to exclude any participant. 86 posts were divided because they contained several arguments. Thus, we collected 493 single argument in total, including 275 for condition A and 218 for condition B.

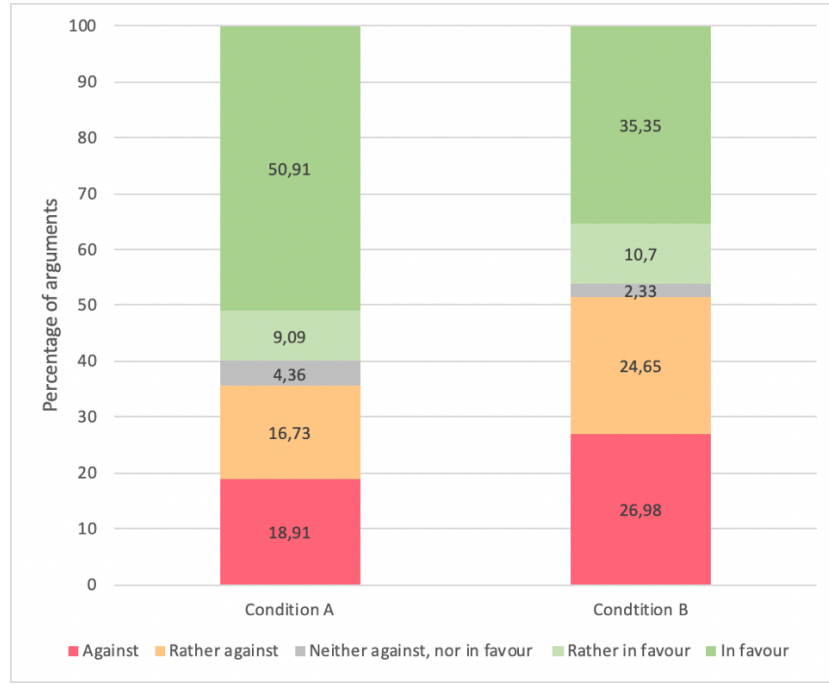


Figure 3. *Percentage of arguments per position*

A chi-squared test on the percentage of arguments per position revealed that there was a significant difference between the two conditions, $X^2(3; N=493) = 16.56, p = .002$. Importantly for us, in both samples there is no overwhelming majority in one sense or the other, since in both conditions there are at least 35% of arguments for each side of the debate. This shows that the urban toll project was, as expected, a controversial project. This increases the ecological validity of our results.

4.2.4. Dependent variables

All the variables were measured based exclusively on the final coding of Coder 2, to make sure that all the arguments were coded based on the same arbitrations so that any observed statistical difference could not be due to a difference of arbitrations between different coders.

Main dependent variable

Our main dependent variable is the probability that a group of n participants will tackle at least x aspect. We note y_i the number of aspects tackled by the group i . We note the probability that y_i would be equal or superior to x for condition A, $P_{n,A}(y_i \geq x)$ and for condition B, $P_{n,B}(y_i \geq x)$. In order to make sure that there was no significant difference in the mean number of aspects assigned to A-participants and B-participants, which would have biased our evaluation of the level of aspect diversity achieved by groups, we performed a t-test to compare

the mean number of assigned aspects per participant for both conditions. There was no significant difference (mean for A-participants: 4.52, mean for B-participants: 4.63, mean difference = - 0.11, 95% confidence interval = [-0.42; 0.21], $t(197) = -0.663$, $p = .508$). This suggests that any difference between A-groups and B-groups cannot be explained by a difference in the level of aspect diversity of the *assigned arguments* but was indeed due to a difference in the level of aspect diversity of the *produced arguments*.

We compared $P_{n.A}(y_i \geq x)$ and $P_{n.B}(y_i \geq x)$ for groups of 2 to 5 participants and for groups of 10, 20, 30, etc., up to 90 participants. We also performed this measure for all possible numbers of aspect (x) going from 1 to 11. Each time, we performed a z-test to compare $P_{n.A}(y_i \geq x)$ and $P_{n.B}(y_i \geq x)$. In most cases, there was no significant difference between the two. Actually, in a large majority of cases, both probabilities were equal either to zero or to 1. This is simply because, for a certain x, if groups are small enough, no group will be able to tackle x aspects, whether or not they are A-groups or B-groups, while if groups are large enough, all the groups will tackle more than x aspect, whether or not they are A-groups or B-groups. In the table 9 below, we present only cases where $p < .250$. In other words, we present only cases where the difference between $P_{n.A}(y_i \geq x)$ and $P_{n.B}(y_i \geq x)$ is at least relatively close to statistical significance.

Table 9. Comparison between $P_{n.A}(y_i \geq x)$ and $P_{n.B}(y_i \geq x)$.

Number of aspect x	Group-size n	$P_{n.A}(y_i \geq x)$	$P_{n.B}(y_i \geq x)$	z value	p	Number of groups
8	2	0.415	0.391	1.205	.228	1000
9	2	0.129	0.111	1.162	.245	1000
	3	0.332	0.302	1.191	.233	1000
	5	0.639	0.588	1.456	.145	1000
10	2	0.041	0.013	3.810	.000***	1000
	3	0.095	0.056	3.174	.001**	1000
	4	0.193	0.122	4.000	.000***	1000
	5	0.289	0.179	5.087	.000***	1000
	10	0.573	0.523	1.510	.131	1000
11	2	0.005	0.001	1.633	.102	1000
	4	0.046	0.019	3.349	.001***	1000
	5	0.093	0.034	5.235	.000***	1000
	10	0.26	0.153	5.265	.000***	1000
	20	0.633	0.554	2.293	.021*	1000
12	4	0.003	0	1.732	.083	1000
	10	0.046	0.011	4.636	.000***	1000
	20	0.247	0.157	4.478	.000***	1000
	30	0.41	0.327	3.057	.002**	1000
	40	0.635	0.499	4.039	.000***	1000
	50	0.772	0.664	2.850	.004**	1000
	60	0.887	0.803	2.043	.041*	1000
	70	0.947	0.89	1.330	.184	1000

* $p < .05$; ** $p < .01$; *** $p < .001$

For all other values of x and n, p was above .250.

It appears that every time there is a significant difference between $P_{n.A}(y_i \geq x)$ and $P_{n.B}(y_i \geq x)$, it is in favour of A-groups. Thus, A-groups have systematically either the same chance or more chance to tackle at least x aspect than B-groups, whatever the value of n and whatever the value of x. For instance, there is 63.5% chance that A-groups of 40 participants will tackle 12 aspects while only 49.9% chance that B-groups of the same size will tackle that number of aspects ($z = 4.039$, $p = .000$). Thus, the results of the group-level variable confirm our main hypothesis: exposure to arguments tends to reduce the level of aspect diversity achieved by groups. We observe statistically significant differences between A-groups and B-groups for groups from 2 participants up to 60 participants. In other words, for groups of 70 participants or more, no significant difference was observed. Thus, B-groups seem to “catch up” with A-groups when they go beyond a certain group-size. However, this “catching up” might be simply due to the fact that, if the pool of all the assigned arguments of the group members tackles on

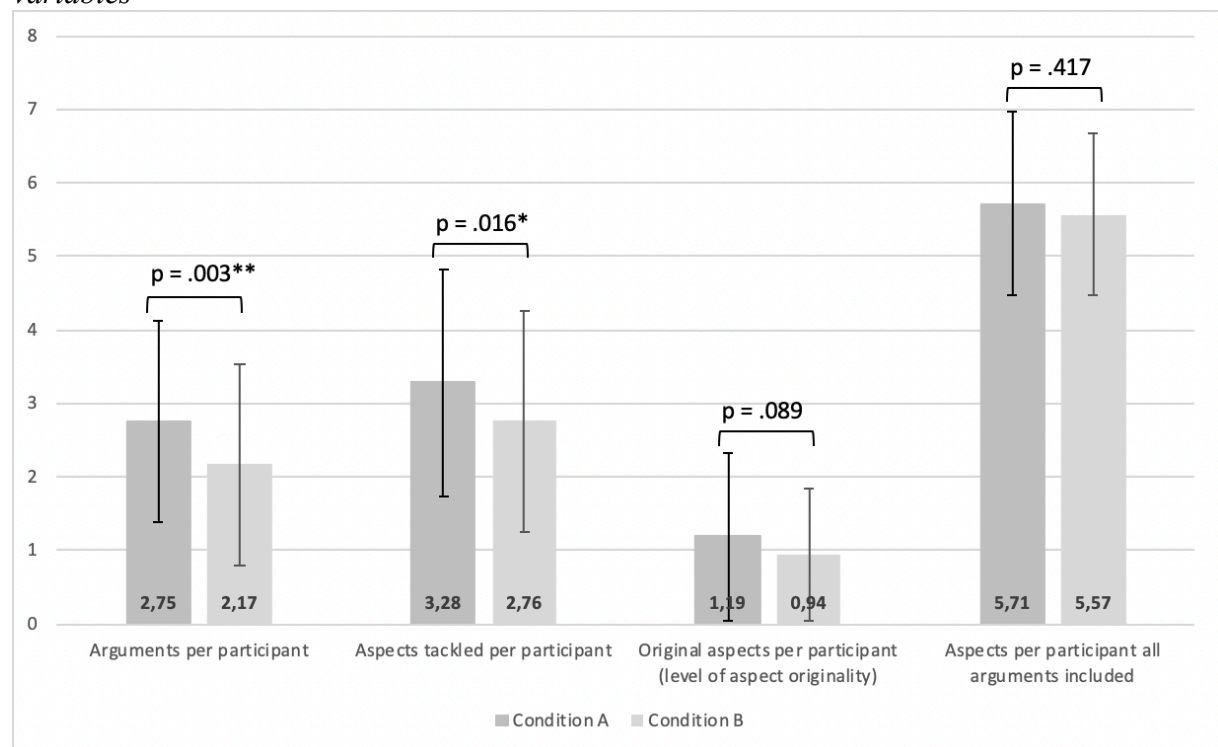
its own most of the existing aspects, then there is no room left for A-participants or B-participant to increase significantly the level of aspect diversity of the group.

As A-participants have been assigned arguments from condition A, there is a chance that the total pool of assigned arguments of some A-groups contains arguments produced by members of the group. In other words, there can be an overlap between the pool of assigned arguments of an A-group and the pool of produced arguments of that group. This could bias to some extent the comparison with B-groups, because in condition B, of course, there can be no such overlap. Thus, to make sure that our results are not significantly biased by this phenomenon, we have measured our group-level variable also when controlling that there was no *overlapping A-group*, i.e. no A-group with an overlap between the pool of assigned arguments and the pool of produced arguments. Results are detailed in appendix VII. We observed a similar pattern in favour of A-groups, which confirms that A-groups are more likely to tackle more aspects than B-groups.

Complementary individual-level variables

We introduced some individual-level variables to obtain complementary data concerning individual behaviour and how it could explain group-level effects.

Figure 4. Comparison between the two conditions regarding complementary individual-level variables



* $p < .05$; ** $p < .01$

The error bars represent the standard deviations.

B-participants tackle on average less aspects than A-participants (difference in means = 0.52, 95% confidence interval = [0.10; 0.95], $t(197) = -2.426$, $p = .016$), and write on average less arguments (mean difference = 0.58, 95% confidence interval = [0.19; 0.96], $t(197) = -2.973$, $p = .003$). Thus, seeing some arguments of other participants does significantly impact the behaviour of participants as far the number of arguments produced and the number of aspects tackled are concerned.

Regarding the level of aspect originality, a difference between B-participants and A-participants was found as a trend (i.e. p is superior to 0.05 but inferior to 0.1) : B-participants tackle a slightly smaller number of original aspects than A-participants (mean difference = 0.25, 95% confidence interval = [-0.04; 0.54], $t(197) = 1.708$, $p = .089$). This trend however does not appear when we consider the mean number of aspects per participant *all arguments included* (mean difference = 0.14, 95% confidence interval = [-0.21; 0.49], $t(197) = 0.813$, $p = .417$). Thus, at the individual level, exposure to some arguments does not seem to impact significantly the extent to which your own arguments will add to the aspect diversity of those arguments.

Those results may seem puzzling considering our results at the group level. At that level, the exposure to others' arguments had a significant impact on the aspect diversity, in other words B-groups tended to tackle less aspects than A-groups. It was natural to expect that this result could be explained by the fact that B-participants tended to tackle a significantly smaller number of original aspects than A-participants. We see two possible explanations for this puzzling situation. The first one is that there is indeed a small difference between A- and B-participants in their disposition to tackle original aspects, even if it does not appear as a significant one in the statistical results, and these small individual differences add up to become statistically significant when analyses are not made at the individual level but at the group level. The second explanation is that even if there is no impact on the number of original aspects tackled, there is an impact concerning *which* aspects are tackled. In other words, B-participants could tackle aspects which are more common (i.e. more frequently tackled) than those tackled by A-participants. In this case, at the group level, there would be more overlap between the original aspects of the different B-participants. This would explain why at the group level, B-participants tend to contribute less to the aspect diversity of the group than A-participants.

In order to investigate the matter a little further and see whether the second explanation was or not likely we compared the percentage of participants tackling each aspect for both conditions. If uncommon aspects were less frequently tackled by B-participants than A-participants, that would bring some support in favour of the plausibility of this explanation.

Figure 5. The percentage of participants tackling each aspect

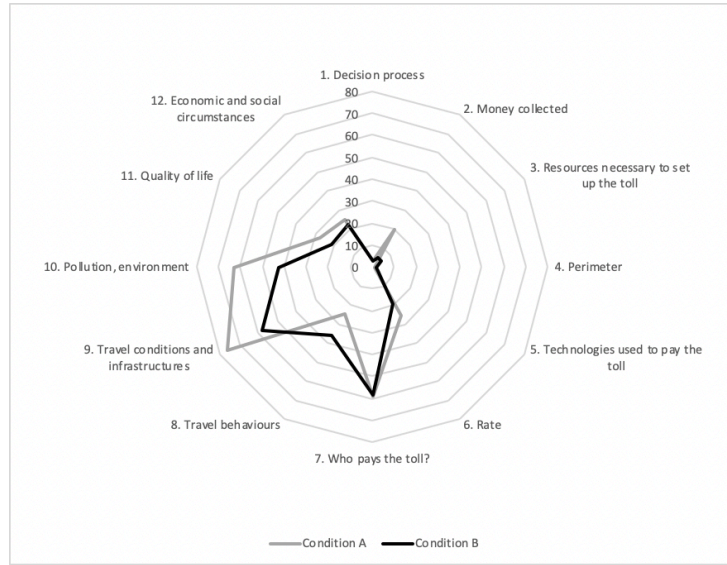


Table 10. The percentage of participants tackling each aspect

Aspect	Condition A	Condition B	Both conditions
1. Decision process	3.0	3.0	3.0
2. Money collected	20.0	5.1	12.6
3. Resources necessary to set up the toll	2.0	5.1	3.5
4. Perimeter	1.0	2.0	1.5
5. Technologies used to pay the toll	1.0	3.0	2.0
6. Rate	26.0	19.2	22.6
7. Who pays the toll?	59.0	58.6	58.8
8. Travel behaviours	25.0	36.4	30.7
9. Travel conditions and infrastructures	76.0	57, 6	66.8
10. Pollution, environment	63.0	42.4	52.8
11. Quality of life	27.0	21.2	24.1
12. Economic and social circumstances	25.0	22.2	23.6

A chi-test did not reveal a significant difference between conditions regarding the prevalence of aspects, $X^2(11; N=199) = 17.671, p = .090$, even if a difference was found as a trend. If we look at the detail, no definitive pattern emerges regarding the probability of B-participants to tackle uncommon aspects. If we consider the aspects tackled by less than 15% of participants in total, we can see that B-participants are less likely to tackle aspects 2 (*Money collected*) but slightly more likely to tackle aspects 3 (*Resources necessary to set up the toll*), 4 (*Perimeter*), 5 (*Technologies used to pay the toll*), and just as likely to tackle aspect 1 (*Decision process*). Of course, this is no definitive proof that B-participants do not tackle, on average, aspects that are more common. However, it is enough to say that although this is possible, it is not the most probable explanation for the observed decrease of aspect diversity at the group-level. Indeed, not only is there no significant difference in the prevalence of aspects between the two conditions and no clear pattern indicating that B-participants are less likely to tackle the most original aspects, but, compared to the other explanation (i.e. that B-participants tackle on average a smaller number of original aspects), the idea that B-participants would tackle as much original aspects as A-participants but that their original aspects would be more common does not make as much sense from a cognitive perspective. Indeed, the fact that B-participants would tackle on average a smaller number of original aspects is coherent with the diverse cognitive processes that we proposed as explanations for our main hypothesis (see section 2.2.1.). Those cognitive processes implied that B-participants would tackle a smaller number

of original aspects than A-participants, because they would focus on aspects tackled by others (due to the will to respond to precise arguments, to normative social influence, or to informational influence) and possibly also because they would make little efforts to find a high number of original aspects due to a downward social comparison effect. On the other hand, in our present state of knowledge, there is no cognitive process, no mechanism or reason that would explain why B-participants might be induced to tackle as much original aspects but ones that are more common than those tackled by A-participants. Thus, even if our analyses do not provide definitive evidence regarding either of the two explanations, the first explanation (that B-participants tend to tackle a smaller number of original aspects) does appear to be the most likely one.

Individual-level variables when controlling for individual characteristics

To make sure that our results were robust, we performed linear regressions to control for the effect of age, gender, level of education and toll-related traits on the individual-level variables. We did not control for the effect of political affiliation because it was less likely to have an impact since urban tolls are not specifically associated with a particular political party, and since environmental concerns were already taken into account. We did not control for the effect of profession either because it was also less likely to have an impact since car use was already taken into account.

Results are detailed in table 11 below. Results appear unchanged: even when controlling for all the demographic and toll-related variables (except for profession and political affiliation), A-participants produce significantly more arguments (regression coefficient for condition B = -0.47 ± 0.20 s.e.m., $t(181) = -2.353$, $p = .020$) and tackle more aspects (regression coefficient for condition B = -0.45 ± 0.23 s.e.m., $t(181) = -1.980$, $p = .049$) than B-participants, while there is no significant difference the number of original aspects per participant (regression coefficient for condition B = -0.19 ± 0.15 s.e.m., $t(181) = -1.240$, $p = .217$) nor in the number of aspects per participant *all arguments included* (regression coefficient for condition B = 0.02 ± 0.18 s.e.m., $t(181) = 0.110$, $p = .912$).

Table 11. Individual-level variables when controlling for age, gender and level of education, for toll related variables, and for all variables (age, gender, education, toll-related variables).

Variable	Regression coefficient	Standard error to the mean	t-value	Degrees of freedom	p
Arguments per participant					
When controlling for age, gender and level of education				189	
Intercept	2.29	0.41	5.625		.000***
Condition B	-0.50	0.20	-2.452		.015*
When controlling for toll-related variables				189	
Intercept	1.71	0.56	3.047		.003**
Condition B	-0.52	0.19	-2.734		.007**
When controlling for all variables				181	
Intercept	1.55	0.71	2.203		.029*
Condition B	-0.47	0.20	-2.353		.020*
Aspects tackled per participant					
When controlling for age, gender and level of education				189	
Intercept	3.55	0.45	7.892		.000***
Condition B	-0.47	0.22	-2.115		.036*
When controlling for toll-related variables				189	
Intercept	2.74	0.64	4.280		.000***
Condition B	-0.49	0.22	-2.260		.025*
When controlling for all variables				181	
Intercept	3.48	0.80	4.368		.000***
Condition B	-0.45	0.23	-1.980		.049*
Original aspects per participant (level of aspect originality)					
When controlling for age, gender and level of education				189	
Intercept	1.13	0.30	3.797		.000***
Condition B	-0.20	0.15	-1.353		.178
When controlling for toll-related variables				189	
Intercept	1.33	0.44	3.037		.003**
Condition B	-0.23	0.15	-1.530		.128
When controlling for all variables				181	
Intercept	1.62	0.53	3.058		.003**
Condition B	-0.19	0.15	-1.240		.217
Aspects per participant all arguments included					
When controlling for age, gender and level of education					.935
Intercept	5.29	0.35	15.204		.000***
Condition B	0.01	0.17	0.082		.935
When controlling for toll-related variables				189	
Intercept	5.26	0.50	10.586		.000***
Condition B	0.06	0.17	0.355		.723
When controlling for all variables				181	
Intercept	4.73	0.62	7.645		.000***
Condition B	0.02	0.18	0.110		.912

* $p < .05$; ** $p < .01$; *** $p < .001$

Conclusion regarding the results of the main study

Our results indicate that exposure to others' arguments significantly impacts the aspects tackled by participants. At the group-level, results confirm our main hypothesis: exposure to others' arguments does tend to decrease the level of aspect diversity achieved by groups, in the sense that it decreases the *probability* for groups to tackle at least a certain number of aspects. However, our experiment does not provide clear answers as to what cognitive processes underlie this phenomenon. When we presented our main hypothesis, we proposed several cognitive mechanisms which could lead to such a result: the will to respond to specific arguments, informational influence, normative social influence, or even a downward social comparison effect. All those mechanisms, though very different, led theoretically to the same result: B-participants were supposed to tackle a smaller number of original aspects. Our results however do not provide definitive evidence that B-participants tackle indeed a smaller number of original aspects on average than A-participants. Though we observed a difference in this direction, it was not high enough to be statistically significant. Nonetheless, as we argued above, we believe that the most probable explanation for our results is the fact that B-participants do indeed tackle less original aspects than A-participants, and that although the difference is too small to be statistically significant at the individual level in our data, it becomes significant at the group-level. Thus, we believe that the decrease in aspect diversity at the group-level might still be due to one or several of the cognitive processes we had presented. Future research will be needed to confirm this and evaluate the respective influence of each of those cognitive processes.

Interestingly, we also discovered that B-participants tackled significantly less aspects than A-participants and produced significantly less arguments. Those results remained unchanged even when controlling for several demographic variables and for toll-related variables, thus they appear very robust. Here too, further research is needed to understand what cognitive processes underlie such phenomena. Indeed, several (compatible) explanations are possible:

- Exposure to others' arguments could decrease one's motivation to produce a high number of arguments or tackle a high number of aspects, for instance because several arguments that more or less coincide with one's general opinion have already been produced, which makes producing more arguments on the same line seem pointless.
- It could incite participants to produce only *new* arguments. In this case, B-participants may produce less arguments simply because they do not repeat common, easily accessible arguments as often as A-participants, because B-participants are aware that those arguments have already been produced by others. Producing less arguments could in this case lead to tackling less aspects.
- Finally, seeing only one argument per position might have biased participants into thinking that other participants generally produced only one argument (and tackled only few aspects). Thus, participants might have been incited by this social norm to produce themselves few arguments and tackle few aspects.

If only the second explanation proved to be the correct, then this phenomenon would not be problematic for consultative platforms. However, if either of the two other explanations proved to be correct, then making others' contributions visible on platform might lead not only to a decrease in aspect diversity but also to a decrease in the number of non-redundant arguments collected.

5. Discussion and recommendations

5.1. Discussion regarding the ecological validity of the study: what can our results tell us about real consultative platforms?

Our results show that exposure to others' argument tends to decrease aspect diversity at the group-level. However, those results were produced in an experimental context, which is of course not equivalent to a real-life consultative platform. Thus, we need to answer the following question: to what extent can our results apply to real platforms? In order to answer this question, we discuss three main limits to the ecological validity of this study.

Firstly, participants did not have the same type of motivation to produce arguments as in a real-life consultative platform. They did not believe that they would be able to influence decision-makers, and they were paid the same amount of money no matter how much arguments they wrote. Besides, on real consultative platforms, citizens self-select to participate, thus only those who have a particular motivation and enough confidence to express themselves end up posting arguments on the platform. Thus, it is probable that our participants were much less motivated than their "real-platform" counterparts to express their opinion, and also probably had to some extent *different* motivations (for instance, the desire to please the experimenters and do what they believed was expected of them). However, all participants were all told that their arguments could be seen by others, and thus believed their arguments had some chance of convincing others. According to the argumentative theory of reasoning, the function of argument production (the function this mechanism evolved to perform) is not to seek the truth but to convince others (Mercier and Sperber, 2011; Mercier, 2016). Thus, believing that your arguments can influence others must be a determinant part of one's motivation to produce arguments. As our participants all had this essential motivation in common with their "real-platform" counterparts, our results can still be considered as representative to some extent of what happens when participants of real-life online platforms engage in argument production.

Secondly, we can't know to what extent the demographic profile of our participant samples is representative of that of their real-platform counterparts. Indeed, to our knowledge, there exists no data concerning the demographic profile of participants on real-life consultative platforms *in general*. Actually, the demographic profile of participants in real-life consultations is likely to vary significantly depending on the subject of the consultation (for instance, a

platform on a retirement reform will probably attract a different type of people than a platform on a reform of school programs). This means that the demographic profile of our participants could be representative to that of some platforms, but more importantly, it means that in order to prove that our results could apply to any (or most) consultative platforms, future studies will need to make sure that they do not depend on some specific demographic feature.

Thirdly, we presented to each B-participant 4 arguments randomly chosen, one for each non-neutral position. This choice on our part was justified by methodological requirements: in order to evaluate strictly the effect of exposure to others' arguments *ceteris paribus*, we needed to expose B-participants to standardized pools of arguments in order to exclude that some uncontrolled parameter might influence the behaviour of the subjects by triggering a specific cognitive mechanism or by intervening on some of the cognitive mechanisms at play. However, on real-life platforms, participants are free to choose how many and what type of arguments they read. As a consequence, maybe on average participants read much more or much less than 4 arguments, maybe they don't read a balanced pool of arguments but read arguments only for one size of the debate, or maybe they choose the arguments they want to read based on other criteria (such as their length, for instance). In any case, the effects of exposure to others' arguments might be different if participants read spontaneously pools of arguments that are very different from the ones we presented them with.

Thus, our experiment is not sufficient to establish that exposure to others' arguments systematically decreases the level of aspect diversity achieved by groups on actual platforms: future research is needed to see whether our results can be replicated in different contexts, with different participants, with different pools of assigned arguments, etc. However, if our results were to be replicated, and if it was confirmed that seeing others' arguments does indeed reduce aspect diversity at the group-level, what would it imply for platform designers? This is the question we try to answer in our following section.

5.2. Recommendations for designing consultative platforms

Importantly, even if it were confirmed that making arguments visible on platforms does reduce aspect diversity, it would *not* imply that all platforms have to make other's arguments invisible. The visibility of others' arguments can have many consequences, and there might be good reasons to make arguments visible even if it means reducing the level of aspect diversity of the group. For instance, making arguments visible could have some positive effect on the number of non-redundant arguments collected and on the quality of arguments (as we will argue in the next section). As a consequence, platform designers must arbitrate between the advantages and disadvantages of making others' arguments visible depending on the specific goals and context of each consultation.

That being said, if making other's arguments visible has indeed a negative impact on aspect diversity, this impact needs to be taken into account carefully by designers. Indeed, there is *a priori* no "easy solution" to solve this problem. For instance, one may think that an easy way around this negative effect would be simply to attract enough participants so that all the aspects would be tackled whatever the design. However, the number of participants that would be required to make this negative effect disappear is not known. Our experiment does not offer any answer on this issue. On this point, in fact, one must be careful in interpreting the results obtained in our experiment, and specially be careful not to over-interpret them. These results show that for groups of 70 participants or more, there is no significant difference in the level of aspect diversity achieved by A-groups and B-groups. However, it does *not* mean that for platforms designers, it would be enough to recruit at least 70 participants to obtain the same level of aspect diversity for platforms with and without visibility of others' arguments. There are two reasons for this.

Firstly, when we observed groups of n participants, we do not observe the level of aspect diversity achieved only by n people, as we take into account also the arguments assigned to each participant of the group. As our pools of assigned arguments have been selected following specific criteria, we cannot measure accurately the number of participants that would be required to produce the equivalent of the pools of assigned arguments. Thus, we cannot say exactly how many participants would be needed to produce a pool of arguments which is equivalent aspect-wise to the pool of assigned arguments *and* to the pool of produced arguments of a group of size n .

Secondly, in most platforms, when other participants' arguments are visible, they are visible from the beginning of the consultation. Participant $n^{\circ}2$ sees the arguments of participant $n^{\circ}1$, participant $n^{\circ}3$ sees the arguments of participants $n^{\circ}1$ and $n^{\circ}2$, etc. This is very different from our experimental design. Indeed, in our experimental design, the pool of assigned arguments all come from condition A. Thus, it is not equivalent to a situation where all participants would have seen others' arguments since the beginning. This is an important point. If it is true that B-groups end up introducing as many aspects as A-groups when they include 70 participants or more, it does not however inform us on what 70 B-participants would have produced if they had each seen others' arguments from the beginning, that is, if participant $n^{\circ}2$ had seen the arguments of participant $n^{\circ}1$, if participant $n^{\circ}3$ had seen the arguments of participants $n^{\circ}1$ and $n^{\circ}2$, etc. Such a situation might create a cascade effect whereby all participants end up focusing more or less on the aspects tackled by the first participants. So, our results don't give the means to evaluate the size of the difference in aspect diversity that would be obtained between a platform where arguments are visible *from the beginning* and one where arguments are invisible.

Thus, our results do not allow to draw any robust conclusion concerning the propensity of platforms on which arguments visible from the beginning to "catch up" with platforms on which arguments are invisible once a certain number of participants has been reached. In other words, our results are not sufficient to prove that "visible arguments" platforms catch up with "invisible arguments" platforms when they attract a certain number of participants, and they certainly do not prove that such a "catching up" happens as soon as these platforms attract 70 participants. Even if "visible-arguments" platforms did catch up eventually, it might well take

groups of much more than 70 participants to do so, and real-life platforms – especially at the local level – do not always attract that many contributions.

Thus, if the negative effect of seeing arguments on aspect diversity was confirmed, there would be *a priori* no easy way around this effect. In that case, what advices could we offer platform designers? First, they should take carefully into consideration the existence of this effect when making their design choices, so that if obtaining a high level of aspect diversity is a priority of the platform, they avoid making others' arguments visible from the beginning. Second, some innovative design solutions could be tested to get around the difficulty. Indeed, visibility and invisibility are not the only design choices possible. There are actually a range of possible options. For instance:

- Others' arguments could be made less easily accessible (one would have to go on a different webpage for instance). This might enable platforms to collect arguments both from participants who did not read any arguments (because they didn't make the effort of actively looking for them) and participants who did.
- A platform could collect a first pool of arguments from participants who did not see any arguments, to make sure to collect a diverse pool of arguments, and then, in a second stage, make this first pool of arguments visible to new participants.
- Others' arguments could be shown to participants only after they have posted a first series of arguments. Thus, participants could first produce contributions without being influenced by others' arguments, and then, if reading others' arguments inspired them, they could add new contributions.

Those are only a few examples. So even if it was definitely proved that seeing arguments decreases aspect diversity, platform designers need not be condemned to choose between favouring aspect diversity and favouring some other advantage of making arguments visible. They could invent and test more complex designs in order to obtain more efficient platforms. However, in order to invent better platforms, platform designers need to know all the relevant information regarding the impact of making arguments visible. This is why, in our last section, we propose some directions for future research on this point.

5.3. Directions for future research: what do platform designers still need to know?

Of course, as said above, the first thing that needs to be investigated is whether or not our findings can be replicated. That being said, if our results were confirmed, if it was an established fact that seeing others' arguments does indeed reduce aspect diversity at the group-level, what would platforms designers still need to know regarding the effects of making arguments visible in order to make informed design decision? In this section, we propose three answers to this question, each one delineating a direction for future researches.

First direction: evaluating the impact of exposure to others' arguments on the other dimension of argument diversity, namely the number of non-redundant arguments collected

There are good reasons to believe that exposure to others' arguments does not only influence aspect diversity but also the number of non-redundant arguments that are collected by the platform. Indeed, studies about idea brainstorming have found that exposure to others' ideas can increase the number of non-redundant ideas produced by groups and individuals (see for instance Yagolkovskiy, 2016; Fink *et al.*, 2012; Leggett Dugosh and Paulus, 2005; Nijstad, Stroebe and Lodewijkx, 2002; Dennis and Valacich, 1999), due to cognitive stimulation and/or social stimulation. It is possible that in the same way seeing others' arguments increases the number of non-redundant arguments produced by groups. Though our results have shown that B-participants produce individually less arguments than A-participants, they could nonetheless produce more original arguments. As a consequence, B-groups could produce a higher number of non-redundant argument than A-groups. What advantage would the fact of collecting more non-redundant arguments bring? A more in-depth exploration of the various aspects tackled. For instance, considering a group of n arguments from A-participant and a group of n arguments from B-participants relative to an aspect, a lower redundancy in the latter means that the n arguments of the B-participants offer more insights into the subject-matter than the n arguments of the A-participants.

If the hypothesis that seeing others' arguments increases the number of non-redundant arguments collected proved to be true, then platforms designers, when choosing whether they will make contributions visible or not, would be faced with a "trade-off" between the number of non-redundant arguments they can collect and the level of aspect diversity they can achieve. Making contributions visible would impact positively one dimension of argument diversity while impacting negatively the other. Knowing whether such a "trade-off" exists is of course essential for designers to make the most efficient design decision based on the specific goals of their platform.

Second direction: evaluating the impact of exposure to others' arguments on argument quality

The goal of consultative platforms is not only to collect a pool of arguments that achieves a high level of diversity, but to collect arguments of a high quality. However, to our knowledge, no study has been made regarding how seeing others' arguments on participatory platforms impacts the quality of the arguments collected. This is why investigating this point would be valuable for platform designers, so that they know whether a "trade-off" exists between argument quality and either dimension of argument diversity.

There are some reasons to believe that exposure to others' arguments could impact positively the quality of the arguments produced by participants. In a real discussion, the *argumentative theory of reasoning* predicts that "the back and forth of dialog enables improvements in argument quality by letting people address successive rounds of

counterarguments” (Mercier, 2016). Indeed, according to the theory, people are cognitively lazy when they produce arguments and they content themselves with arguments that may convince others, without anticipating any counterargument to their claim. But if people are confronted with counterarguments, they are able to evaluate them objectively, which improves their reflection, and they become able to produce arguments that are not invalidated by those counterarguments. Thus, addressing counterarguments creates an improvement in the quality of people’s arguments.

In a platform, being confronted with other people’s arguments could have the same effect, although with a smaller effect size than in a discussion. Though the platform does not allow for successive rounds of counterargument, addressing one round of counterarguments could still lead to some improvement in argument quality compared to addressing none. However, the context of a platform is of course very different from a classical deliberation process (and the argumentative theory does not make predictions regarding this particular context). Thus, it is possible that seeing counterarguments arguments *in a platform* would not have the same effect as being confronted with counterarguments in a real-life discussion.

Third direction: evaluating how many arguments and what type of arguments people read when they have the choice

As said above, in real life platforms, participants are free to choose how many and what type of arguments they read. To our knowledge, no study has yet been made regarding how many arguments people read on average on consultative platforms, nor regarding what type of argument they choose to read (the firsts that appear? the ones that coincide with their own view? the shorter ones?). This is a problem, since the effect of the visibility of arguments may vary depending on what type of arguments participant choose to read.

For instance, there are evidence that people prefer being exposed to information that support their view, and that it may lead them to practice “selective exposure”, i.e. to seek out only information congruent with their views (for a review of literature on this point, see Hallsworth *et al.*, 2018). This may lead participants of consultative platforms to read only arguments congruent with their views. Such a selective exposure might have several consequences, among which, possibly, an influence on the tendency of participants to focus on the same aspects as previous arguments. Indeed, some evidence suggest that participants might be more inclined to tackle the same aspects as comments which are congruent with their own views (McInnis *et al.*, 2018).

Thus, in order to better evaluate the impact of making arguments visible on a platform, and to better design future experiments on this point, one would also need to investigate those questions.

Conclusion

The aim of the work presented in this thesis has been to make a first contribution to a new line of research: the impact of design features of consultative platforms on argument diversity. The experiment we made has delivered a series of significative results relative to the design feature we chose to test (the exposure to arguments of other participants). It showed that this feature has a negative impact on one of the two dimensions of diversity, namely aspect diversity. As we have already stressed, these results need to be confirmed and many researches need to be done before one can understand better what underlies this phenomenon, and more generally how this feature, as well as others, impact argument diversity relative to all its facets.

That being said, what conclusions could platform designers draw from our results if they were confirmed? Not that one should never make arguments visible on a platform. This would be an oversimplification of matters: design choices in consultative platforms must fulfil numerous requirements, and favouring aspect diversity is only one among many desirable outcomes that platform designers may wish to achieve. The lesson should not be either that it is enough to attract a certain (reasonable) number of participants in order to counter the negative effect in question: there is for the moment no proof that this would work. Thus, the take-home message should rather be that platform designers must carefully weight all the advantages and disadvantages of making arguments visible when they make their decision, considering the context and the specific goals of the consultation they are about to launch. And, importantly, they should not get trapped in the binary choice of making the arguments either visible from start or never visible. Numerous possibilities exist beside those two choices. The best course of action may be to start exploring these possibilities.

Bibliography

ADEME (Agence de l'environnement et de la maîtrise de l'énergie) (2016), *Etat de l'art sur les péages urbain*. Available at: <https://www.ademe.fr/etat-lart-peages-urbains> (Accessed: 31 March 2021).

Aitamurto, T. (2016) 'Crowdsourced democratic deliberation in open policymaking: Definition, promises, challenges', in Crowdsourced democratic deliberation in open policymaking: Definition, promises, challenges. International Reports on Socio-Informatics (IRSI), Proceedings of the CSCW 2016–Workshop: Toward a Typology of Participation in Crowdsourcing, pp. 67–78.

Aitamurto, T. and Chen, K. (2017) 'The value of crowdsourcing in public policymaking: epistemic, democratic and economic value', *The theory and practice of legislation*, 5(1), pp. 55–72.

Aitamurto, T. and Landemore, H. (2016) 'Crowdsourced deliberation: The case of the law on off-road traffic in Finland', *Policy & Internet*, 8(2), pp. 174–196.

Aitamurto, T. and Landemore, H. E. (2015) 'Five design principles for crowdsourced policymaking: Assessing the case of crowdsourced off-road traffic law in Finland', *Journal of Social Media for Organizations*, 2(1), pp. 1–19.

Brabham, D. C. and Guth, K. L. (2017) 'The deliberative politics of the consultative layer: Participation hopes and communication as design values of civic tech founders', *Journal of Communication*, 67(4), pp. 445–475.

Dennis, A. R., Minas, R. K. and Williams, M. L. (2019) 'Creativity in computer-mediated virtual groups', *The Oxford handbook of group creativity and innovation*, p. 253-269.

Dennis, A. R. and Valacich, J. S. (1999) 'Research Note. Electronic Brainstorming: Illusions and Patterns of Productivity', *Information Systems Research*, 10(4), pp. 375–377. doi: 10.1287/isre.10.4.375.

Eliasson, J. (2016) 'Is congestion pricing fair? Consumer and citizen perspectives on equity effects', *Transport policy*, 52, pp. 1–15.

Eliasson, J. and Jonsson, L. (2011) 'The unexpected “yes”: Explanatory factors behind the positive attitudes to congestion charges in Stockholm', *Transport Policy*, 18(4), pp. 636–647.

Falco, E. and Kleinhans, R. (2019) 'Digital participatory platforms for co-production in urban development: A systematic review', *Crowdsourcing: Concepts, Methodologies, Tools, and Applications*, pp. 663–690.

Fink, A., Koschutnig, K., Benedek, M., Reishofer, G., Ischebeck, A., Weiss, E.M., and Ebner, F. (2012) 'Stimulating creativity via the exposure to other people's ideas', *Human Brain Mapping*, 33(11), pp. 2603–2610. doi: 10.1002/hbm.21387.

- Gallupe, R. B., Dennis, A.R., Cooper, W.H., Valacich, J.S., Bastianutti, L.M., and Nunamaker, J.F. (1992) 'Electronic Brainstorming and Group Size', *The Academy of Management Journal*, 35(2), pp. 350–369. doi: 10.2307/256377.
- Hallsworth, M., Egan, M., Rutter, J., and McCrae, J. (2018) 'Behavioural government. Using behavioural science to improve how governments make decisions'. The Behavioural Insights Team. Available at <https://www.bi.team/publications/behavioural-government/>
- Hong, L. and Page, S. E. (2001) 'Problem solving by heterogeneous agents', *Journal of economic theory*, 97(1), pp. 123–163.
- Hong, L. and Page, S. E. (2004) 'Groups of diverse problem solvers can outperform groups of high-ability problem solvers', *Proceedings of the National Academy of Sciences*, 101(46), pp. 16385–16389.
- Jaensirisak, S., Wardman, M. and May, A. D. (2005) 'Explaining variations in public acceptability of road pricing schemes', *Journal of Transport Economics and Policy (JTEP)*, 39(2), pp. 127–154.
- Koch, G., Füller, J. and Brunswicker, S. (2011) 'Online crowdsourcing in the public sector: how to design open government platforms', in *International Conference on Online Communities and Social Computing*. Springer, pp. 203–212.
- Kopp, P. and Prud'homme, R. (2010) 'The economics of urban tolls: Lessons from the Stockholm case', *International Journal of Transport Economics / Rivista internazionale di economia dei trasporti*, 37(2), pp. 195–221.
- Landemore, H. (2011) 'Democratic Reason: The Mechanisms of Collective Intelligence in Politics', *Collective Wisdom: Principles and Mechanisms*. doi: 10.1017/CBO9780511846427.012.
- Landemore, H. (2013), *Democratic reason: politics, collective intelligence, and the rule of the many*. Princeton, N.J. Oxford: Princeton University Press.
- Leggett Dugosh, K. and Paulus, P. B. (2005) 'Cognitive and social comparison processes in brainstorming', *Journal of Experimental Social Psychology*, 41(3), pp. 313–320. doi: 10.1016/j.jesp.2004.05.009.
- Lu, L., Yuan, Y. C. and McLeod, P. L. (2012) 'Twenty-five years of hidden profiles in group decision making: A meta-analysis', *Personality and Social Psychology Review*, 16(1), pp. 54–75.
- Manin, B. (2004) 'Délibération et discussion', *Revue suisse de science politique*, 10(4), pp. 180–192.
- McInnis, B., Cosley, D., Baumer, E., Leshed, G. (2018) 'Effects of Comment Curation and Opposition on Coherence in Online Policy Discussion', in *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, pp. 347–358.
- Mercier, H. (2011) 'Looking for Arguments', *Argumentation*, 26, pp. 1–20. doi: 10.1007/s10503-011-9256-1.

- Mercier, H. (2016) 'The Argumentative Theory: Predictions and Empirical Evidence', *Trends in Cognitive Sciences*, 20(9), pp. 689–700. doi: 10.1016/j.tics.2016.07.001.
- Mercier, H., Boudry, M., Paglieri, F., and Trouche, E. (2017) 'Natural-born arguers: Teaching how to make the best of our reasoning abilities', *Educational Psychologist*, 52(1), pp. 1–16.
- Mercier, H. and Sperber, D. (2011) 'Why do humans reason? Arguments for an argumentative theory', *Behavioral and Brain Sciences*, 34(2), pp. 57–74. doi: 10.1017/S0140525X10000968.
- Nelimarkka, M., Nonnecke, B., Krishnan, S., Aitumurto, T., Catterson, D., Crittenden, C., et al. (2014). 'Comparing Three Online Civic Engagement Platforms using the Spectrum of Public Participation'. *UC Berkeley: Center for Information Technology Research in the Interest of Society (CITRIS)*. Retrieved from <https://escholarship.org/uc/item/0bz755bj>
- Neuendorf, K. A. (2017) *The content analysis guidebook*. 2nd edition. Los Angeles, CA: SAGE Publications, Inc., chapters 1 to 6
- Nijstad, B, Bechtoldt, M & Choi, H-S (2019), 'Information processing, motivation, and group creativity'. in PB Paulus & BA Nijstad (eds), *The Oxford handbook of group creativity and innovation*. Oxford library of psychology, Oxford University Press, New York, pp. 87-102.
- Nijstad, B. A., Stroebe, W. and Lodewijkx, H. F. M. (2002) 'Cognitive stimulation and interference in groups: Exposure effects in an idea generation task', *Journal of Experimental Social Psychology*, 38(6), pp. 535–544. doi: 10.1016/S0022-1031(02)00500-0.
- Ottaviani, M. and Sørensen, P. (2001) 'Information Aggregation in Debate: Who should Speak First?', *Journal of Public Economics*, 81, pp. 393–421. doi: 10.1016/S0047-2727(00)00119-5.
- Page, S. E. (2008), 'Unpacking the toolbox', in *The Difference : How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton, Oxford: Princeton University Press. Available at: <https://univ-scholarvox-com.acces-distant.sciencespo.fr/book/45003634> (Accessed: 26 November 2020).
- Paulus, P. B. and Dzindolet, M. T. (1993) 'Social influence processes in group brainstorming.', *Journal of Personality and Social Psychology*, 64(4), p. 575.
- Paulus, P. B. and Kenworthy, J. B. (2019) 'Effective brainstorming', in PB Paulus & BA Nijstad (eds), *The Oxford handbook of group creativity and innovation*. Oxford library of psychology, Oxford University Press, New York, pp. 287–306.
- R. Farina, C., Epstein, D., Heidt, J., Newhart, M.J. (2013) 'Regulation Room: Getting “more, better” civic participation in complex government policymaking', *Transforming Government: People, Process and Policy*, 7(4), pp. 501–516. doi: 10.1108/TG-02-2013-0005.
- Raux, C. and Souche, S. (2004) 'Comment améliorer l'acceptation du péage urbain?', in *Entretiens Jacques Cartier. XVIIèmes conférence, Colloque 5 Transports en commun et transports routiers urbains: qui doit payer? = Who must pay for urban road and transit services?*, 7-8 oct. 2004, Montréal-Québec-Sherbrooke, pp. 11-p.
- Simon, J., Bass, T., Boelman, V., Mulgan, G. (2017) 'Digital Democracy. The Tools Transforming Political Engagement', Nesta. Available at:

<https://www.nesta.org.uk/report/digital-democracy-the-tools-transforming-political-engagement/>

Souche-Le Corvec, S., Raux, C., Eliasson, J., Hamilton, C., Brundell-Freij, K., Kiiskilä, K., Tervonen, J. (2016) ‘Predicting the results of a referendum on urban road pricing in France: “the cry of Cassandra”?’’, *European Transport Research Review*, 8(2), p. 15.

Sunstein, C. R. and Hastie, R. (2015), *Wiser: Getting beyond groupthink to make groups smarter*. Boston, Mass.: Harvard Business Press.

Taeihagh, A. (2017) ‘Crowdsourcing: a new tool for policy-making?’, *Policy Sciences*, 50(4), pp. 629–647.

Toplak, M. E. and Stanovich, K. E. (2003) ‘Associations between myside bias on an informal reasoning task and amount of post-secondary education’, *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 17(7), pp. 851–860.

Towne, W. B. and Herbsleb, J. D. (2012) ‘Design considerations for online deliberation systems’, *Journal of Information Technology & Politics*, 9(1), pp. 97–115.

Wittenbaum, G. M., Hubbell, A. P. and Zuckerman, C. (1999) ‘Mutual enhancement: Toward an understanding of the collective preference for shared information’, *Journal of Personality and Social Psychology*, 77(5), pp. 967–978. doi: 10.1037/0022-3514.77.5.967.

Wittenbaum, G. M. and Park, E. S. (2001) ‘The Collective Preference for Shared Information’, *Current Directions in Psychological Science*, 10(2), pp. 70–73. doi: 10.1111/1467-8721.00118.

Yagolkovskiy, S. R. (2016) ‘Stimulation of Individual Creativity in Electronic Brainstorming: Cognitive and Social Aspects’, *Social Behavior and Personality: an international journal*, 44(5), p. 761.

Ziegler, R., Diehl, M. and Zijlstra, G. (2000) ‘Idea Production in Nominal and Virtual Groups: Does Computer-Mediated Communication Improve Group Brainstorming?’, *Group Processes & Intergroup Relations*, 3(2), pp. 141–158. doi: 10.1177/1368430200032003.

Online sources

Cambridge dictionary (no date). *Definition of argument*. [online]. Available at: <https://dictionary.cambridge.org/dictionary/english/argument> [accessed March 2021]

French Government (2019), *Vers un revenu universel d’activité*. [online]. Available at: <https://www.consultation-rua.gouv.fr/> [accessed March 2021]

French Government (2020), *Consultation citoyenne sur le Service Civique*. [online]. Available at: <https://consultation.service-civique.gouv.fr/> [accessed March 2021]

Australian Government (no date), *Consultation Hub*. [online]. Available at: <https://consultations.health.gov.au/> [accessed March 2021]

Scottish Government (no date), *Consultation Hub*. [online]. Available at: <https://consult.gov.scot/> [accessed March 2021]

Delib (no date), *Who uses delib*. [online]. Available at: https://www.delib.net/who_uses_delib [accessed March 2021]

Cap Collectif (no date), *Réalisations*. [online]. Available at: <https://cap-collectif.com/realisations-2/> [accessed March 2021]

Appendix

Appendix I. Instructions

1. General description of urban tolls given to participants

French

*Mettre en place un péage urbain signifie rendre payant l'accès automobile à un ou plusieurs quartiers d'une ville. Autrement dit, quand il y a un péage urbain autour d'un quartier, **les véhicules motorisés (voiture, camion, etc.) qui veulent circuler dans ce quartier doivent payer**. Le péage ne s'applique pas aux transports en commun.*

***Le tarif peut varier ou être fixe.** Par exemple, le tarif peut changer en fonction de l'heure ou du jour de la semaine. Il peut aussi changer en fonction du type de véhicule, ou encore en fonction de la raison pour laquelle la personne veut circuler dans le quartier (travail, loisir, etc.). Le tarif peut aussi changer en fonction d'autres facteurs.*

*Le **but** du péage urbain est de diminuer le nombre de véhicules dans le quartier et/ou de récolter de l'argent.*

English translation

*Setting up an urban toll means charging for the access of motor vehicles to one or several neighbourhoods of a city. In other words, when there is an urban toll around a neighbourhood, **motor vehicles (cars, trucks, etc.) which want to drive in this neighbourhood have to pay**. Public transports do not have to pay the toll.*

***The rate can vary or it can be a fixed price.** For instance, the rate can vary depending on the hour of the day or the day of the week. It can also vary depending on the type of vehicle, or depending on the motive for which the person wants to drive in the neighbourhood (work, leisure activities, etc.). The rate can also vary depending on other factors.*

*The **goal** of the urban toll is to reduce the number of motor vehicles inside the neighbourhood and/or to collect money.*

2. Description of the fictional project of urban toll given to participants

French

Imaginez la situation suivante :

Le maire de votre ville ou de la ville que vous visitez le plus souvent pense à mettre en place un péage urbain. Le péage urbain serait autour du centre-ville. L'objectif du péage serait de limiter la pollution et les bouchons. Il s'agirait d'un forfait journalier : les véhicules qui veulent circuler dans le centre-ville devraient payer le péage une fois par jour. Après avoir payé le

English translation

Imagine the following situation :

The mayor of your city, or of the city you visit the most frequently, thinks about setting up an urban toll. The urban toll would be around the city centre. The goal of the toll would be to limit pollution and congestion. It would consist in a daily rate: vehicles wanting to drive in the city centre should pay the toll once a day. After having paid the toll, vehicles could drive inside the city

péage, les véhicules pourraient circuler dans le centre-ville, en sortir et y revenir librement pendant la journée. Les résidents du centre-ville paieraient un tarif réduit, toutes les autres personnes paieraient plein tarif. L'argent récolté par le péage serait utilisé pour améliorer le réseau de bus, de tramway et de métro.

Le maire souhaite consulter la population **sur la pertinence du projet et les différentes options possibles**. Il lance donc une plateforme participative en ligne. **Vous décidez d'exprimer votre opinion sur cette plateforme.**

centre, drive out of it and come back freely during the day. City centre residents would pay a reduced rate, all the other people would pay full rate. The money collected through the toll would be used to improve the bus, tram and underground networks.

The mayor wishes to consult the population about **the relevance of the project and the different possible options**. As a consequence, he launches an online participatory platform. **You decide to express your opinion on this platform.**

3. Instructions given to participants relative to the arguments they had to produce

French

Vous pouvez poster **un ou plusieurs** arguments au sujet du projet de péage urbain.
Un argument = une raison pour soutenir le projet **ou** une raison pour s'opposer au projet.
Si vous trouvez que le projet a à la fois des aspects positifs et des aspects négatifs, vous pouvez poster à la fois des arguments pour et des arguments contre.
Vous pourrez relire la présentation des péages urbains et la description du projet de péage si besoin.
Séparez bien vos arguments (**un seul argument par post**). Écrivez des phrases complètes.
Vos arguments pourront être montrés aux prochains participants.

English translation

You can post **one or several** arguments concerning the project of urban toll.
An argument = a reason to support the project **or** a reason to oppose the project. If you think that the project has both positive features and negative features, you can post both arguments for the project and arguments against it.
You will be able to read again the presentation of urban tolls and the description of the project, if you need to.
Distinguish carefully your arguments from one another (**only one argument per post**). Write complete sentences.
Your arguments may be shown to future participants.

4. Instructions given to participants from condition B concerning the arguments they had to read

French

Avant de poster votre (vos) argument(s) sur la plateforme, vous lisez quelques arguments des

English translation

Before posting your argument(s) on the platform, you read some arguments written by people who

gens qui se sont exprimés avant vous sur la plateforme.

Instructions :

*Les arguments que vous allez voir ont été choisis au hasard parmi ceux proposés par les précédents participants à cette expérience. Certains participants ont produit plusieurs arguments¹, mais tous les arguments d'un même participant ne sont pas nécessairement présents ici. **Lisez tous les arguments.** Après les avoir lus, vous pourrez vous-même poster un ou plusieurs arguments au sujet du projet de péage urbain.*

expressed their opinion on the platform before you.

Instructions:

*The arguments you are going to see have been chosen randomly among those written by the previous participants of this experiment. Some participants have produced several arguments, but all the arguments of one participant are not necessarily present here. **Read all the arguments.** After having read them, you will yourself be able to post one or several arguments concerning the urban toll project.*

Appendix II. Attention check questions about urban tolls

French

1. *S'il y a un péage urbain, vous payez quand*
 - ☐ vous utilisez un véhicule motorisé
 - ☐ vous utilisez n'importe quel véhicule, motorisé ou non
 - ☐ vous utilisez une voiture, un vélo ou un tramway
 - ☐ Vous utilisez un tramway, un métro ou un bus
2. *S'il y a un péage urbain, vous payez quand :*
 - ☐ Vous conduisez sur une route financée par l'Etat
 - ☐ Vous conduisez à l'intérieur d'un quartier d'une ville
 - ☐ Vous conduisez n'importe où avec une voiture polluante
 - ☐ Vous n'aidez pas à financer les transports publics
3. *Dans le projet proposé, l'argent récolté par le péage urbain doit être utilisé pour :*
 - ☐ Aider les gens à acheter des voitures électriques
 - ☐ Améliorer les bus, métro et tramways
 - ☐ Améliorer les espaces de parking
 - ☐ Rien n'est dit sur l'usage de l'argent.
4. *Dans le projet proposé, certaines personnes payent un tarif réduit :*
 - ☐ Les personnes qui rendent visite à quelqu'un habitant dans le centre-ville
 - ☐ Les personnes qui reçoivent des allocations sociales (RSA, chômage, etc.)
 - ☐ Les personnes qui ne font que traverser le centre-ville
 - ☐ Les personnes qui habitent dans le centre-ville
5. *Dans le projet proposé, le péage devait être payé :*
 - ☐ Toutes les heures
 - ☐ Tous les jours
 - ☐ Toutes les semaines
 - ☐ Tous les mois

English translation

1. *If there is an urban toll, you pay when*
 - ☐ you use a motor vehicle
 - ☐ you use any vehicle, whether it is a motor vehicle or not
 - ☐ you use a car, a bicycle, or a tramway
 - ☐ you use a tramway, the underground, or a bus
2. *If there is an urban toll, you pay when :*
 - ☐ You drive on a publicly financed road
 - ☐ You drive inside a neighbourhood of a city
 - ☐ You drive anywhere with a polluting car
 - ☐ You do not help to finance public transports
3. *In the proposed urban toll project, the money collected through the toll shall be used to :*
 - ☐ Help people by electric cars
 - ☐ Improve the bus, tram and underground networks
 - ☐ Improve the parking spots
 - ☐ Nothing is said on the use of the money.
4. *In the proposed urban toll project, some people pay a reduced rate:*
 - ☐ People who come to visit someone who lives inside the city-centre
 - ☐ People who benefit from social welfare
 - ☐ People who only drive through the city-centre
 - ☐ People who live inside the city-centre
5. *In the proposed urban toll project, the toll had to be paid:*
 - ☐ Every hour
 - ☐ Every day
 - ☐ Every week
 - ☐ Every month

Note : for questions 1 to 4, answers were presented in a random order to participants.

Appendix III. Arguments failing the relevance check

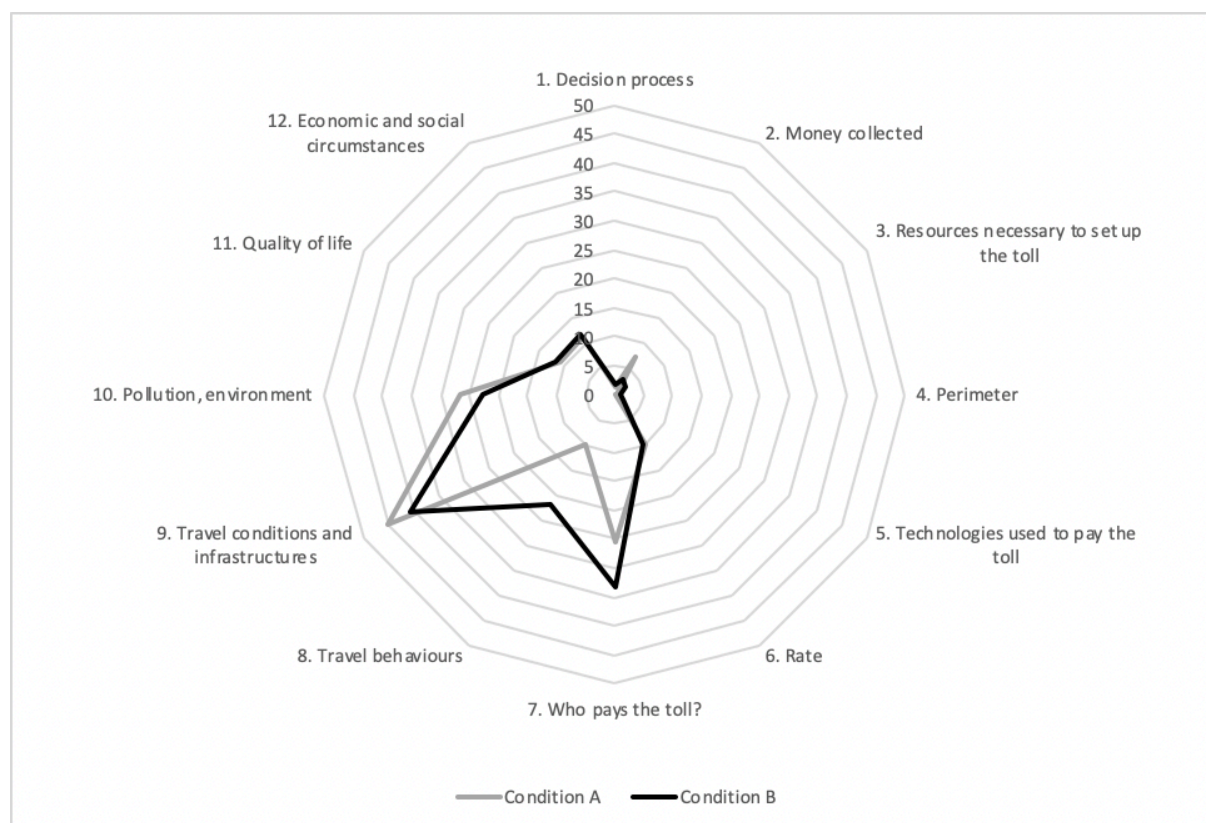
- *“Par contre sur nationale et départementale les poids étranger eux doivent payer pour l'entretien ou le feroutage”*
- *“L'idée d'installer un péage autour d'un centre-ville est absurde. Les gens pour le contrer facilement en arrivant pied.”*

Appendix IV. List of aspects

1. No precise aspect
2. Decision process
3. Money collected
4. Resources necessary (time, money) for the setting up and maintenance of the urban toll
5. Perimeter of the toll
6. Methods, technologies and infrastructures used to carry out and control the payment of the toll
7. The rate of the toll and its variations depending on hour of the day, day of the week, or else
8. Who pays the toll?
9. Travel-related behaviours
10. Driving conditions, or other travel- related circumstances; travel-related infrastructures
11. Air pollution, environment, health (not transportation-related)
12. Quality of life, town planning (not transportation-related)
13. Economic and social circumstances (not transportation-related)
14. Other aspect

Appendix V. Percentage of arguments tackling each aspect

Figure A1. *Percentage of arguments tackling each aspect*



Appendix VI. Coding reliability for the main study before correction of the careless mistakes

Table A1. Reliability measures for the coding of aspects in condition A before correction of the careless mistakes

Aspect	Number of agreements	Number of disagreements	Cohen's kappa
0. No precise aspect	1	1	0.66
1. Decision process	1	0	1
2. Money collected	9	0	1
3. Resources necessary to set up the toll	0	1	0
4. Perimeter	0	0	1
5. Technologies used to pay the toll	1	0	1
6. Rate	4	3	0.70
7. Who pays the toll?	11	4	0.8
8. Travel behaviours	6	8	0.51
9. Travel conditions and infrastructures	17	3	0.88
10. Pollution, environment	12	1	0.95
11. Quality of life	7	3	0.79
12. Economic and social circumstances	4	3	0.70
13. Other aspect	0	0	1
Total	73	27	/
Mean (per aspect)	5.2	1.9	/

Table A2. Reliability measures for the coding of aspects in the final coding before correction of the careless mistakes

Aspect	Number of agreements	Number of disagreements	Cohen's kappa
0. No precise aspect	0	0	1
1. Decision process	2	0	1
2. Money collected	4	3	0.72
3. Resources necessary to set up the toll	3	1	0.85
4. Perimeter	0	0	1
5. Technologies used to pay the toll	2	0	1
6. Rate	12	1	0.96
7. Who pays the toll?	37	5	0.91
8. Travel behaviours	23	16	0.67
9. Travel conditions and infrastructures	50	14	0.79
10. Pollution, environment	28	3	0.93
11. Quality of life	19	3	0.91
12. Economic and social circumstances	12	5	0.81
13. Other aspect	0	0	1
Total	192	49	/
Mean (per aspect)	13.7	3.5	/

Appendix VII. Comparison between $P_{n.A}(y_i \geq x)$ and $P_{n.B}(y_i \geq x)$ without overlapping A-groups

We could not obtain enough groups of more than 10 participants without creating overlapping A-groups. Thus, we calculated $P_{n.A}(y_i \geq x)$ and $P_{n.B}(y_i \geq x)$ only for groups of 2, 3, 4, 5 and 10 participants. Moreover, for those group-sizes, it was not possible to obtain 1000 A-groups without overlapping groups. Thus, the following measures are based on less than 1000 A-groups.

Table A3. Comparison between $P_{n.A}(y_i \geq x)$ and $P_{n.B}(y_i \geq x)$ without overlapping A-groups

Number of aspect x	Group-size n	$P_{n.A}(y_i \geq x)$	$P_{n.B}(y_i \geq x)$	z value	p	Number of A-groups	Number of B-groups
8	2	0.423	0.381	1.460	.144	914	1000
9	2	0.132	0.111	1.337	.181	914	1000
	5	0.644	0.588	1.230	.219	430	1000
10	2	0.039	0.013	3.523	.000***	914	1000
	3	0.093	0.056	2.809	.005**	771	1000
	4	0.190	0.122	3.299	.001***	620	1000
	5	0.281	0.179	3.547	.000***	430	1000
11	2	0.005	0.001	1.691	.091	914	1000
	4	0.053	0.019	3.343	.001***	620	1000
	5	0.088	0.034	3.513	.000***	430	1000
12	4	0.005	0	1.732	.083	620	1000

* $p < .05$; ** $p < .01$; *** $p < .001$

For all other values of x and n, p was above .250.

Appendix VIII. Regression results for individual-level variables

Table A4. Regression results for the mean number of arguments per participant

	Regression coefficient	Standard error to the mean	t-value	df	p
When controlling for age, gender and level of education				189	
Intercept	2.29	0.41	5.625		.000***
Condition B	-0.50	0.20	-2.452		.015*
Age	0.005	0.011	0.436		.663
Gender – male	0.095	0.204	0.465		.643
Education - <i>Brevet des collèges</i> (secondary school diploma)	-0.382	1.407	-0.272		.786
Education - <i>CAP, BEP</i> , other same-level diploma	-1.038	1.401	-0.741		.460
Education - Master's degree	0.401	0.291	1.381		.169
Education - PhD	0.436	0.491	0.888		.376
Education - <i>Licence, BTS, DUT</i> , other <i>BAC+2</i> qualification	0.058	0.299	0.193		.847
Education - Does not wish to say	2.574	1.407	1.830		.069
When controlling for toll-related variables				189	
Intercept	1.71	0.56	3.047		.003**
Condition B	-0.52	0.19	-2.734		.007**
Car use – at least twice a month	0.853	0.278	3.069		.002**
Car use – at least twice a week	0.373	0.255	1.460		.146
Car use – every day	0.660	0.272	2.428		.016*
Interest in environmental issues	-0.003	0.178	-0.015		.988
Importance of travelling in an environmental-friendly way	0.284	0.147	1.927		.056
Worry about environmental issues	-0.145	0.185	-0.783		.435
Did you know what an urban toll was before this experiment – more or less	0.101	0.243	0.416		.678
Did you know what an urban toll was before this experiment –Yes	0.229	0.245	0.937		.350
When controlling for all variables				181	
Intercept	1.55	0.71	2.203		.029*
Condition B	-0.47	0.20	-2.353		.020*
Car use – at least twice a month	0.904	0.293	3.085		.002**
Car use – at least twice a week	0.327	0.261	1.253		.212
Car use – every day	0.610	0.281	2.173		.031*
Interest in environmental issues	-0.019	0.180	-0.103		.918

Importance of travelling in an environmental-friendly way	0.293	0.150	1.956	.052
Worry about environmental issues	-0.132	0.189	-0.698	.486
Did you know what an urban toll was before this experiment – more or less	0.098	0.251	0.393	.695
Did you know what an urban toll was before this experiment –Yes	0.196	0.265	0.739	.461
Age	0.000	0.011	0.023	.981
Gender –male	0.119	0.212	0.563	.574
Education - <i>Brevet des collèges</i> (secondary school diploma)	-0.990	1.411	-0.702	.484
Education - <i>CAP, BEP</i> , other same-level diploma	-2.159	1.416	-1.524	.129
Education - Master's degree	0.162	0.294	0.550	.583
Education - PhD	0.244	0.498	0.490	.625
Education - <i>Licence, BTS, DUT</i> , other <i>BAC+2</i> qualification	-0.068	0.298	-0.227	.820
Education - Does not wish to say	1.883	1.407	1.338	.183

* $p < .05$; ** $p < .01$; *** $p < .001$

Table A5. Regression results for the mean number of aspects tackled per participant

	Regression coefficient	Standard error to the mean	t-value	df	p
When controlling for age, gender and level of education				189	
Intercept	3.55	0.45	7.892		.000***
Condition B	-0.47	0.22	-2.115		.036*
Age	-0.017	0.012	-1.418		.158
Gender – male	-0.111	0.225	-0.492		.623
Education - <i>Brevet des collèges</i> (secondary school diploma)	-1.215	1.552	-0.783		.435
Education - <i>CAP, BEP</i> , other same-level diploma	-1.423	1.546	-0.920		.359
Education - Master's degree	0.361	0.321	1.127		.261
Education - PhD	0.030	0.542	0.056		.956
Education - <i>Licence, BTS, DUT</i> , other <i>BAC+2</i> qualification	0.307	0.330	0.931		.353
Education - Does not wish to say	2.941	1.552	1.895		.060
When controlling for toll-related variables				189	
Intercept	2.74	0.64	4.280		.000***
Condition B	-0.49	0.22	-2.260		.025*

Car use – at least twice a month	0.498	0.318	1.567	.119
Car use – at least twice a week	0.514	0.292	1.763	.080
Car use – every day	0.387	0.311	1.248	.214
Interest in environmental issues	-0.028	0.203	-0.138	.890
Importance of travelling in an environmental-friendly way	0.031	0.168	0.184	.854
Worry about environmental issues	0.042	0.211	0.199	.842
Did you know what an urban toll was before this experiment – more or less	0.149	0.278	0.536	.593
Did you know what an urban toll was before this experiment –Yes	0.177	0.280	0.633	.528
When controlling for all variables			181	
Intercept	3.48	0.80	4.368	.000***
Condition B	-0.45	0.23	-1.980	.049*
Car use – at least twice a month	0.465	0.331	1.406	.161
Car use – at least twice a week	0.574	0.295	1.406	.053
Car use – every day	0.524	0.317	1.655	.010*
Interest in environmental issues	-0.059	0.203	-0.289	.773
Importance of travelling in an environmental-friendly way	0.057	0.169	0.334	.739
Worry about environmental issues	-0.034	0.213	-0.158	.875
Did you know what an urban toll was before this experiment – more or less	0.193	0.283	0.681	.497
Did you know what an urban toll was before this experiment –Yes	0.389	0.299	1.302	.195
Age	-0.023	0.013	-1.788	.076
Gender – male	-0.200	0.240	-0.835	.405
Education - <i>Brevet des collèges</i> (secondary school diploma)	-1.591	1.593	-0.999	.319
Education - <i>CAP, BEP</i> , other same-level diploma	-1.827	1.599	-1.143	.255
Education - Master's degree	0.211	0.332	0.636	.526
Education - PhD	-0.260	0.562	-0.462	.645
Education - <i>Licence, BTS, DUT</i> , other <i>BAC+2</i> qualification	0.264	0.337	0.782	.435
Education - Does not wish to say	2.710	1.589	1.705	.090

* $p < .05$; ** $p < .01$; *** $p < .001$

Table A6. Regression results for the mean number of original aspects per participant (level of aspect diversity)

	Regression coefficient	Standard error to the mean	t-value	df	p
When controlling for age, gender and level of education				189	
Intercept	1.13	0.30	3.797		.000***
Condition B	-0.20	0.15	-1.353		.178
Age	0.002	0.008	0.302		.763
Gender – male	-0.151	0.149	-1.017		.310
Education - <i>Brevet des collèges</i> (secondary school diploma)	-1.174	1.027	-1.144		.254
Education - <i>CAP, BEP</i> , other same-level diploma	-0.852	1.022	-0.834		.405
Education - Master's degree	0.023	0.212	0.110		.913
Education - PhD	-0.388	0.358	-1.083		.280
Education - <i>Licence, BTS, DUT</i> , other <i>BAC+2</i> qualification	0.191	0.218	0.875		.383
Education - Does not wish to say	3.804	1.026	3.707		.000***
When controlling for toll-related variables				189	
Intercept	1.33	0.44	3.037		.003**
Condition B	-0.23	0.15	-1.530		.128
Car use – at least twice a month	0.188	0.217	0.865		.389
Car use – at least twice a week	-0.108	0.199	-0.541		.589
Car use – every day	0.111	0.212	0.525		.600
Interest in environmental issues	0.116	0.139	0.837		.404
Importance of travelling in an environmental-friendly way	-0.103	0.115	-0.900		.369
Worry about environmental issues	-0.093	0.144	-0.645		.520
Did you know what an urban toll was before this experiment – more or less	0.135	0.190	0.709		.479
Did you know what an urban toll was before this experiment –Yes	-0.015	0.191	-0.079		.937
When controlling for all variables				181	
Intercept	1.62	0.53	3.058		.003**
Condition B	-0.19	0.15	-1.240		.217
Car use – at least twice a month	0.099	0.221	0.450		.653
Car use – at least twice a week	-0.027	0.197	-0.138		.890
Car use – every day	0.192	0.211	0.910		.364
Interest in environmental issues	0.064	0.136	0.476		.635
Importance of travelling in an environmental-friendly way	-0.078	0.113	-0.693		.489
Worry about environmental issues	-0.139	0.142	-0.980		.328

Did you know what an urban toll was before this experiment – more or less	0.089	0.189	0.472	.638
Did you know what an urban toll was before this experiment –Yes	0.073	0.199	0.366	.715
Age	-0.001	0.009	-0.080	.936
Gender –male	-0.222	0.160	-1.389	.167
Education - <i>Brevet des collèges</i> (secondary school diploma)	-1.339	1.062	-1.261	.209
Education - <i>CAP, BEP</i> , other same-level diploma	-0.739	1.066	-0.694	.489
Education - Master's degree	0.046	0.221	0.210	.834
Education - PhD	-0.368	0.375	-0.983	.327
Education - <i>Licence, BTS, DUT</i> , other <i>BAC+2</i> qualification	0.212	0.225	0.943	.347
Education - Does not wish to say	3.756	1.059	3.546	.000***

* $p < .05$; ** $p < .01$; *** $p < .001$

Table A7. Regression results for the mean number of aspects per participant all arguments included

	Regression coefficient	Standard error to the mean	t-value	df	p
When controlling for age, gender and level of education					.935
Intercept	5.29	0.35	15.204		.000***
Condition B	0.01	0.17	0.082		.935
Age	0.003	0.009	0.296		.768
Gender – male	0.273	0.174	1.567		.119
Education - <i>Brevet des collèges</i> (secondary school diploma)	-0.342	1.202	-0.285		.776
Education - <i>CAP, BEP</i> , other same-level diploma	1.337	1.197	1.117		.265
Education - Master's degree	0.070	0.248	0.281		.779
Education - PhD	0.319	0.420	0.761		.448
Education - <i>Licence, BTS, DUT</i> , other <i>BAC+2</i> qualification	0.062	0.255	0.244		.808
Education - Does not wish to say	-2.367	1.202	-1.970		.050
When controlling for toll-related variables					189
Intercept	5.26	0.50	10.586		.000***
Condition B	0.06	0.17	0.355		.723
Car use – at least twice a month	0.05	0.25	0.223		.824

Car use – at least twice a week	0.41	0.23	1.833	.068
Car use – every day	-0.04	0.24	-0.186	.853
Interest in environmental issues	-0.01	0.16	-0.042	.967
Importance of travelling in an environmental-friendly way	0.03	0.13	0.242	.809
Worry about environmental issues	0.00	0.16	0.010	.992
Did you know what an urban toll was before this experiment – more or less	0.16	0.22	0.754	.452
Did you know what an urban toll was before this experiment –Yes	0.18	0.22	0.840	.402
When controlling for all variables			181	
Intercept	4.73	0.62	7.645	.000***
Condition B	0.02	0.18	0.110	.912
Car use – at least twice a month	0.171	0.257	0.666	.506
Car use – at least twice a week	0.366	0.229	1.597	.112
Car use – every day	-0.124	0.246	-0.505	.614
Interest in environmental issues	0.013	0.158	0.081	.935
Importance of travelling in an environmental-friendly way	0.016	0.132	0.118	.906
Worry about environmental issues	0.069	0.166	0.418	.677
Did you know what an urban toll was before this experiment – more or less	0.181	0.220	0.822	.412
Did you know what an urban toll was before this experiment –Yes	0.056	0.232	0.242	.809
Age	0.005	0.010	0.497	.620
Gender – male	0.337	0.186	1.809	.072
Education - <i>Brevet des collèges</i> (secondary school diploma)	-0.469	1.238	-0.379	.705
Education - <i>CAP, BEP</i> , other same-level diploma	1.113	1.243	0.896	.371
Education - Master's degree	0.010	0.258	0.040	.968
Education - PhD	0.217	0.437	0.497	.620
Education - <i>Licence, BTS, DUT</i> , other <i>BAC+2</i> qualification	-0.000	0.262	-0.001	.999
Education - Does not wish to say	-2.624	1.235	-2.125	.035*

* $p < .05$; ** $p < .01$; *** $p < .001$

This series presents the Master's theses in Public Policy and in European Affairs of the Sciences Po School of Public Affairs. It aims to promote high-standard research master's theses, relying on interdisciplinary analyses and leading to evidence-based policy recommendations.

How to favour argument diversity on online consultative platforms? An experiment on the effect of exposure to other participants' arguments on the diversity of aspects tackled

Sophie de Rouilhan

Abstract

On online consultative platforms - a type of digital democracy tool where citizens are asked to put forth arguments relative to a public policy project - what matters is not only the quality of the posted arguments but also their diversity. Especially, the pool of arguments collected is expected to tackle all the aspects of the project under consideration. The number of aspects tackled by the pool of arguments can be influenced by the design of the platform, especially by whether or not the arguments of other participants are made visible. Thus, in this thesis, we try to answer the following question: how does the visibility of the arguments of previous participants on consultative platforms impact the number of aspects tackled by the collected pool of arguments? Existing literature in psychology suggests that it should lead to a decrease in the level of aspect diversity achieved by groups, because of the will to respond to others' arguments, informational influence, normative social influence, or possibly a downward social comparison effect. To test this hypothesis, we designed an online experiment, whereby we asked participants to produce arguments as if they were on a consultative platform. Depending on the condition they were put in, they could either see four arguments or none. We compared, using a resampling technique, the probability that same-sized groups from the two conditions would tackle at least a certain number of aspects. The results show that exposure to arguments does tend to decrease the probability that groups achieve a high level of aspect diversity. Finally, we discuss directions for future research and possible implications of those results for platform designers.

Key words

digital democracy, e-participation, consultative platforms, argumentation, argument diversity, semantic categories