The Social Dilemma of Autonomous Vehicles

Article in Science - June 2016
DOI: 10.1126/science.aaf2654

CITATIONS 8
READS 6,386

3 authors, including:

Azim Shariff
University of California, Irvine
49 PUBLICATIONS 1,974 CITATIONS
SEE PROFILE

Iyad Rahwan
Massachusetts Institute of Technology
127 PUBLICATIONS 2,743 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Complex systems and other things View project

All content following this page was uploaded by Azim Shariff on 23 June 2016.
The user has requested enhancement of the downloaded file. All in-text references underlined in blue are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.
to K). The same result was observed with trunk neural crest transplanted with the early cranial-specific factors (n = 0/6 embryos). Reprogrammed trunk neural crest, however, acquired chondrogenic potential and formed ectopic cartilage nodules (n = 4/7 embryos) (Fig. 4, L to O) in the proximal jaw. Thus, introducing components of the cranial-specific transcriptional circuit is sufficient to reprogram trunk neural crest cells and to drive them to adopt an additional cartilaginous fate. These results definitively show that the cranial-specific regulatory circuit (Fig. 3J) we have defined confers chondrogenic potential to the trunk neural crest.

The development and differentiation of neural crest cells are controlled by a complex gene regulatory network, composed of transcription factors, signaling molecules, and epigenetic modifiers (12, 13). We have expanded the known cranial neural crest gene regulatory network by identifying transcriptional interactions specific to the cranial crest and absent from other subpopulations. By linking anterior identity in the gastrula to the expression of drivers of chondrocytic differentiation, we have identified a cranial-specific circuit (Fig. 3J) that endows the neural crest with its potential to differentiate into the craniofacial skeleton of vertebrates. Our results highlight how transcriptional circuits can be rewired to alter progenitor cell identity and fate during embryonic development.

REFERENCES AND NOTES

1. N. Le Douarin, The Neural Crest (Cambridge Univ. Press, 1982).


ETHICS

The social dilemma of autonomous vehicles

Jean-François Bonnefon,1 Azim Shariff,2* Idy Rahnwan3†

Autonomous vehicles (AVs) should reduce traffic accidents, but they will sometimes have to choose between two evils, such as running over pedestrians or sacrificing themselves and their passenger to save the pedestrians. Defining the algorithms that will help AVs make these moral decisions is a formidable challenge. We found that participants in six Amazon Mechanical Turk studies approved of utilitarian AVs (that is, AVs that sacrifice their passengers for the greater good) and would like others to buy them, but they would themselves prefer to ride in AVs that protect their passengers at all costs. The study participants disapprove of enforcing utilitarian regulations for AVs and would be less willing to buy such an AV. Accordingly, regulating for utilitarian algorithms may paradoxically increase casualties by postponing the adoption of a safer technology.

The year 2007 saw the completion of the first benchmark test for autonomous driving in realistic urban environments (1, 2). Since then, autonomous vehicles (AVs) such as Google’s self-driving car covered thousands of miles of real-road driving (3). AVs have the potential to benefit the world by increasing traffic efficiency (4), reducing pollution (5), and eliminating up to 90% of traffic accidents (6). Not all crashes will be avoided, though, and some crashes will require AVs to make difficult ethical decisions in cases that involve unavoidable harm (7). For example, the AV may avoid harming several pedestrians by swerving and sacrificing a passerby, or the AV may be faced with the choice of sacrificing its own passenger to save one or more pedestrians (Fig. 1). Although these scenarios appear unlikely, even low-probability events are bound to occur with millions of AVs on the road. Moreover, even if these situations were never to arise, AV programming must still include rule decisions about what to do in such hypothetical situations. Thus, these types of decisions need be made well before AVs become a global commodity. Distributing harm is a decision that is universally considered to fall within the moral domain (8, 9). Accordingly, the algorithms that control AVs will need to embed moral principles guiding their decisions in situations of unavoidable harm (10). Manufacturers and regulators will need to accomplish three potentially incompatible objectives: being consistent, not causing public outrage, and not discouraging buyers.

However, pursuing these objectives may lead to moral inconsistencies. Consider, for example, the case displayed in Fig. 1A, and assume that

SCIENCE sciencemag.org 24 JUNE 2016 • VOL 352 ISSUE 6293 1573

14. California, Irvine, Irvine, CA 92697, USA.

3The Media Lab, Massachusetts Institute of Technology, CSAIL, Cambridge, Massachusetts 02139, USA.

†Toulouse School of Economics, Toulouse, France.

*To whom correspondence should be addressed. E-mail: irahwan@mit.edu

†Corresponding author.
Detailed statistical results for all studies are provided in the supplementary materials (tables S1 to S8). Overall, participants strongly agreed that it would be more moral for AVs to sacrifice their own passengers when this sacrifice would save a greater number of lives overall.

In study one (n = 182 participants), 76% of participants thought that it would be more moral for AVs to sacrifice one passenger rather than kill 10 pedestrians [with a 95% confidence interval (CI) of 69 to 82]. These same participants were later asked to rate which was the most moral way to program AVs, on a scale from 0 (protect the passenger at all costs) to 100 (minimize the number of casualties). They overwhelmingly expressed a moral preference for utilitarian AVs programmed to minimize the number of casualties (median = 85) (Fig. 2A). However, participants were less certain that AVs would be programmed in a utilitarian manner (67% thought so, with a median rating of 70). Thus, participants were not worried about AVs being too utilitarian, as often portrayed in science-fiction works. If anything, they imagined future AVs as being less utilitarian than they should be.

In study two (n = 451 participants), participants were presented with dilemmas that varied the number of pedestrians’ lives that could be saved, from 1 to 100. Participants did not think that AVs should sacrifice their passenger when only one pedestrian could be saved (with an average approval rate of 20%), but their moral approval increased with the number of lives that could be saved (P < 0.001), up to approval rates consistent with the 76% observed in study one (Fig. 2B).

Participants’ approval of passenger sacrifice was even more robust to treatments in which they had to imagine themselves and another person, particularly a family member, in the AV (study three, n = 259 participants). Imagining that a family member was in the AV negatively affected the morality of the sacrifice, as compared with imagining oneself alone in the AV (P = 0.003). But even in that strongly aversive situation, the morality of the sacrifice was still rated above the midpoint of the scale, with a 95% CI of 54 to 66 (Fig. 3A).

Still, study three presents the first hint of a social dilemma. On a scale of 1 to 100, respondents were asked to indicate how likely they would be to buy an AV programmed to minimize casualties (which would, in these circumstances, sacrifice them and their co-riding family member), as well as how likely they would be to buy an AV programmed to prioritize protecting its passengers, even if it meant killing 10 or 20 pedestrians. Although the reported likelihood of buying an AV was low even for the self-protective option (median = 50), respondents indicated a significantly lower likelihood (P < 0.001) of buying the AV when they imagined the situation in which they and their family member would be sacrificed for the greater good (median = 19). In other words, even though participants still agreed that utilitarian AVs were the most moral, they preferred the self-protective model for themselves.

Study four (n = 267 participants) offers another demonstration of this phenomenon. Participants were given 100 points to allocate between different types of algorithms, to indicate (i) how moral the algorithms were, (ii) how comfortable participants were for other AVs to be programmed in a given manner, and (iii) how likely participants would be to buy an AV programmed in a given manner. For one of the algorithms, the AV would always swerve when it was about to run over people on the road. Figure 3B shows the points allocated to the AV equipped with this algorithm, in three situations: (i) when it swerved into a pedestrian to save 10 people, (ii) when it killed its own passenger to save 10 people, and (iii) when it swerved into a pedestrian to save just one other pedestrian. The algorithm that swerved into one to save 10 always received many points, and the algorithm that swerved into one to save one always received few points. The algorithm that would kill its passenger to save 10 presented a hybrid profile. Like the high-valued algorithm, it received high marks for morality (median budget share = 50) and was considered a good algorithm for other people to have (median budget share = 50). But in terms of purchase intention,
it received significantly fewer points than the high-valued algorithm ($P < 0.001$) and was, in fact, closer to the low-valued algorithms (median budget share = 33). Once more, it appears that people praise utilitarian, self-sacrificing AVs and welcome them on the road, without actually wanting to buy one for themselves.

This is the classic signature of a social dilemma, in which everyone has a temptation to free-ride instead of adopting the behavior that would lead to the best global outcome. One typical solution in this case is for regulators to enforce the behavior leading to the best global outcome. Indeed, there are many similar societal examples involving trade-off of harm by people and governments (15–17). For example, some citizens object to regulations that require children to be immunized before starting school. In this case, the parental decision-makers choose to minimize the perceived risk of harm to their child while increasing the risk to others. Likewise, recognition of the threats of environmental degradation have prompted government regulations aimed at curtailing harmful behaviors for the greater good. But would people approve of government regulations imposing utilitarian algorithms in AVs, and would they be more likely to buy AVs under such regulations?

In study five ($n = 376$ participants), we asked participants about their attitudes toward legally enforcing utilitarian sacrifices. Participants considered scenarios in which either a human driver or a control algorithm had an opportunity to self-sacrifice to save 1 or 10 pedestrians (Fig. 3C). As usual, the perceived morality of the sacrifice was high and about the same whether the sacrifice was performed by a human or by an algorithm (median = 70). When we inquired whether participants would agree to see such moral sacrifices legally enforced, their agreement was higher for algorithms than for human drivers ($P < 0.002$), but the average agreement still remained below the midpoint of the 0 to 100 scale in each scenario. Agreement was highest in the scenario in which AVs were allowed on the market, few people praise utilitarian, self-sacrificing AVs and indicate that they would be less likely to purchase an AV under such regulations [(C) and (D)].

Finally, in study six ($n = 393$ participants), we asked participants specifically about their likelihood of purchasing the AVs whose algorithms had been regulated by the government. Participants were presented with scenarios in which they were riding alone, with an unspecified family member, or with their child. As in the previous studies, the scenarios depicted a situation in which the algorithm that controlled the AV could sacrifice its passengers to minimize casualties on the road.

Participants indicated whether it was the duty of the government to enforce regulations that would minimize the casualties in such circumstances, whether they would consider the purchase of an AV under such regulations, and whether they would consider purchasing an AV under no such regulations. As shown in Fig. 3D, people were reluctant to accept governmental regulation of utilitarian AVs. Even in the most favorable condition, when participants imagined only themselves being sacrificed to save 10 pedestrians, the 95% CI for whether people thought it was appropriate for the government to regulate this sacrifice was only 4 to 32. Finally, participants were much less likely to consider purchasing an AV with such regulation than without ($P < 0.001$). The median expressed likelihood of purchasing an unregulated AV was 59, compared with 21 for purchasing a regulated AV. This is a huge gap from a statistical perspective, but it must be understood as reflecting the state of public sentiment at the very beginning of a new public issue and is thus not guaranteed to persist.

Three groups may be able to decide how AVs handle ethical dilemmas: the consumers who buy the AVs; the manufacturers that program the AVs; and the government, which may regulate the kind of programming manufacturers can offer and consumers can select. Although manufacturers may engage in advertising and lobbying to influence consumer preferences and government regulations, a critical collective problem consists of deciding whether governments should regulate the moral algorithms that manufacturers offer to consumers.

Our findings suggest that regulation for AVs may be necessary but also counterproductive. Moral algorithms for AVs create a social dilemma (18, 19). Although people tend to agree that everyone would be better off if AVs were utilitarian (in the sense of minimizing the number of casualties on the road), these same people have a personal incentive to ride in AVs that will protect them at all costs. Accordingly, if both self-protective and utilitarian AVs were allowed on the market, few people would be willing to ride in utilitarian AVs, even though they would prefer others to do so. Regulation may provide a solution to this problem, but regulators will be faced with two difficulties: First, most people seem to disapprove of a regulation that would enforce utilitarian AVs. Second—and a more serious problem—our results suggest that such regulation could substantially delay the adoption of AVs, which means that the...
lives saved by making AVs utilitarian may be outnumbered by the deaths caused by delaying the adoption of AVs altogether. Thus, car-makers and regulators alike should be considering solutions to these obstacles.

Moral algorithms for AVs will need to tackle more intricate decisions than those considered in our surveys. For example, our scenarios did not feature any uncertainty about decision outcomes, but a collective discussion about moral algorithms will need to encompass the concepts of expected risk, expected value, and blame assignment. Is it acceptable for an AV to avoid a motorcycle by swerving into a wall, considering that the probability of survival is greater for the passenger of the AV than for the rider of the motorcycle? Should AVs account for the ages of passengers and pedestrians (20)? If a manufacturer offers different versions of its moral algorithm, and a buyer knowingly chose one of them, is the buyer to blame for the harmful consequences of the algorithm’s decisions? Such liability considerations will need to accompany existing discussions of regulation (21), and we hope that psychological studies inspired by our own will be able to inform this discussion.

Figuring out how to build ethical autonomous machines is one of the sternest challenges in artificial intelligence today (22). As we are about to endow millions of vehicles with autonomy, a serious consideration of algorithmic morality has never been more urgent. Our data-driven approach highlights how the field of experimental ethics can provide key insights into the moral, cultural, and legal standards that people expect from autonomous driving algorithms. For the time being, there seems to be no easy way to design algorithms that would reconcile moral values and personal self-interest—let alone account for different cultures with various moral attitudes regarding life-life trade-offs (23)—but public opinion and social pressure may very well shift as this conversation progresses.

REFERENCES AND NOTES

ACKNOWLEDGMENTS
J.-F.B. gratefully acknowledges support through the Agence Nationale de la Recherche—Laboratoires d’Excellence Institute for Advanced Study in Toulouse. This research was supported by internal funds from the University of Oregon to A.S. U.R. is grateful for financial support from R. Hoffman. Data files have been uploaded as supplementary materials.

SUPPLEMENTARY MATERIALS
www.sciencemag.org/content/352/6293/suppl/DC1
Materials and Methods
Supplementary Text
Fig. S1
Tables S1 to S8
Data Files S1 to S6
15 January 2016, accepted 21 April 2016
10.1126/science.aaf2654

PROSTATE DEVELOPMENT

Identification of an Nkx3.1-G9a-UTY transcriptional regulatory network that controls prostate differentiation

Aditya Dutta,1* Clémentine Le Magnen,1 Antonina Mitrofanova,2† Xuesong Ouyang,3‡ Andrea Califano,4 Cory Abate-Shen5‡§

The Nkx3.1 homeobox gene plays essential roles in prostate differentiation and prostate cancer. We show that loss of function of Nkx3.1 in mouse prostate results in down-regulation of genes that are essential for prostate differentiation, as well as up-regulation of genes that are not normally expressed in prostate. Conversely, gain of function of Nkx3.1 in an otherwise fully differentiated nonprostatic mouse epithelium (seminal vesicle) is sufficient for respecification to prostate in renal grafts in vivo. In human prostate cells, these activities require the interaction of Nkx3.1 with the G9a histone methyltransferase via the homeodomain and are mediated by activation of target genes such as UTY (KDMD6a), the male-specific paralog of UTX (KDMD6a). We propose that an Nkx3.1-G9a-UTY transcriptional regulatory network is essential for prostate differentiation, and we speculate that disruption of such a network predisposes to prostate cancer.

Among the tissues of the male urogenital system, the prostate and seminal vesicle are secretory organs that develop in close proximity under the influence of androgens (fig. S1A) (1, 2). However, the prostate develops from the urogenital sinus, an endodermal derivative, whereas the seminal vesicle develops from the Wolffian duct, a mesodermal derivative. Among genes that distinguish prostate and seminal vesicle, the Nkx3.1 homeobox gene is among the earliest expressed in the presumptively prostatic epithelium during development, and its expression in adults is primarily restricted to prostatic luminal cells (3, 4), which are secretory cells that are the major target of prostate neoplasia (4, 5). Accordingly, in mouse models, loss of function of Nkx3.1 results in impaired prostate differentiation and defects in luminal stem cells, as well as predisposes to prostate cancer (3, 4).

Analyses of expression profiles from Nkx3.1 wild-type (Nkx3.1+/+) and Nkx3.1 mutant (Nkx3.1−/−) prostates revealed down-regulation of genes associated with prostate differentiation, such as FoxA1 (Forhead Box A1), Psen (Probasin), HoxB13, and Tmprss2 (Transmembrane Protease, Serine 2), as well as luminal cells (cytokeratins 8 and 18), and up-regulation of basal cell markers (cytokeratins 5 and 6) (Fig. 1A, fig. S2A, and database SI) (6). Surprisingly, Nkx3.1−/− versus Nkx3.1+/+ prostates display up-regulation of genes that are expressed, albeit not exclusively, in seminal vesicle, namely

*Departments of Medicine and Urology, Institute of Cancer Genetics, Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY 10032, USA.
†Department of Systems Biology, Columbia University Medical Center, New York, NY 10032, USA.
‡Departments of Urology, Columbia University Medical Center, New York, NY 10032, USA.
§Departments of Systems Biology, Biomedical Informatics, and Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Institute of Cancer Genetics, Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY 10032, USA.
¶Departments of Systems Biology, Biomedical Informatics, and Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Institute of Cancer Genetics, Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY 10032, USA.
©Departments of Urology, Institute of Cancer Genetics, Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY 10032, USA.
**Department of Systems Biology, Biomedical Informatics, and Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Institute of Cancer Genetics, Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY 10032, USA.
††Departments of Urology, Institute of Cancer Genetics, Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY 10032, USA.
‡‡Departments of Systems Biology, Biomedical Informatics, and Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Institute of Cancer Genetics, Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY 10032, USA.
§§Corresponding author. Email: cabateshen@columbia.edu
Editor's Summary

Codes of conduct in autonomous vehicles

When it becomes possible to program decision-making based on moral principles into machines, will self-interest or the public good predominate? In a series of surveys, Bonnefon et al. found that even though participants approve of autonomous vehicles that might sacrifice passengers to save others, respondents would prefer not to ride in such vehicles (see the Perspective by Greene). Respondents would also not approve regulations mandating self-sacrifice, and such regulations would make them less willing to buy an autonomous vehicle.

Science, this issue p. 1573; see also p. 1514